2008

Synthetic data methods for disclosure limitation

Jennifer C. Huckett
Iowa State University

NOTE TO USERS

This reproduction is the best copy available.

Synthetic data methods for disclosure limitation

by

Jennifer C. Huckett

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Michael D. Larsen, Major Professor
Petrutza C. Caragea
Craig Gundersen
Stephen B. Vardeman
Cindy L. Yu

Iowa State University
Ames, Iowa
2008

DEDICATION

This dissertation is dedicated to my mom Sara and my sisters Anna and Nora, to my brother Ben, my dad Steve and my step-dad Brian, to my aunt Leah and my grandma Pat, and to my friends Monica, Joy, Brie, Sarah, Casi, and Eden, and to my wonderful Wayne. My efforts during graduate school, to do this research and to complete this work, have been supported with the constant love that these people have given me. Each of them is inspiring to me and each of them contributes to who I am and how I have done this. Through laughing with me and urging me to live well, I have accomplished a doctorate degree with you.

Mike Larsen has been the best advisor I have had as a student and I thank him for his support. The time and energy and thoughtfulness he has given toward my research and development as a Statistician are invaluable, as is his tolerance for down to the wire revising. I hope to one day fill an adivising or mentoring role this well.

Joanne Wendleberger urged me to study Statistics seven years ago and I am happy with where I am because of it. Professionally, I've learned a lot about Statistics and about learning itself. Personally, I've gotten a little tougher and am better for it. Thank you Joanne for continuing to give me advice, I value it very much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION AND BACKGROUND

In this dissertation we address the general problem faced by statistical agencies to provide data to researchers, policy-makers, and other legitimate data users while upholding respondents' confidentiality and protecting privacy. In the field of Statistics, this type of research belongs to the field of confidentiality protection and disclosure limitation. In Section 1.1, we put this problem into context and discuss the importance of confidentiality protection and disclosure limitation. Approaches to disclosure limitation are discussed in Section 1.2. This leads to considering synthetic data methods as an approach to disclosure limitation. An introduction to synthetic data and a review of current methods used to produce synthetic data are presented in Section 1.3. In Section 1.4 we propose using quantile regression as an improved method to produce synthetic data. In Section 1.5, we describe our proposed use of hot deck imputation and rank swapping to complement quantile regression. The motivation for addressing this problem and an application at the Iowa Department of Revenue are introduced in Section 1.6. In Section 1.7, we introduce work done to study this method at the U. S. Census Bureau. An outline of this dissertation is provided in Section 1.8.

1.1 The importance of confidentiality protection and disclosure limitation

Federal statistical agencies exist in the United States and other countries to inform the public on matters that affect the welfare of the people, both individually and collectively (Duncan, Jabine, and de Wolf 1993). Each of over 70 statistical agencies in the United States was founded in response to specific needs for data about critical areas in public policy (Duncan, et al. 1993). Policy makers and researchers use data products to study inflation and unemployment rates and crime and healthcare statistics allowing them to manage the economy and inform public debate (Doyle, Lane, Theeuwes, and Zayatz 2001). Policy makers make decisions about fund allocation, monitor social programs, investigate potential effects of new legislation, and validate or extend theoretical social science models (Fienberg and

Willenborg 1998). Such research ultimately impacts the competitiveness of the economy by providing policy makers with complete and accurate information, thereby enabling them to make decisions (Doyle, et al. 2001).

Statistical agencies aim to collect and disseminate quality data and information to policymakers and researchers. Agencies face several challenges to ensure the collection and subsequent dissemination of quality data. There are limiting factors. In order to provide quality data to users, an agency must first collect quality data. Data collection poses challenges outside the scope of this dissertation. In this work, we assume quality data have been collected. We address the challenge faced by agencies to provide quality information to users under the constraint that agencies are bound by legal and internal obligations to protect the confidentiality of individuals and organizations, or respondents, from which data are collected. This challenge is reflected in most agencies' mission or policy statements. Such statements usually include explicit goals to collect and provide quality data and to honor and protect privacy and confidentiality. Examples can be found on agencies' websites.

Examples and quotations from three agencies' websites are provided below. The mission statement of the U.S. Census Bureau states:

> *The Census Bureau serves as the leading source of quality data about the nation's people and economy. We honor privacy, protect confidentiality, share our expertise globally, and conduct our work openly. We are guided on this mission by our strong and capable workforce, our readiness to innovate, and our abiding commitment to our customers (U.S. Census Bureau 2008).*

The guiding principle stated in the Mission and Policy on Microdata Dissemination at the National Center for Health Statistics (NCHS) states:

> *NCHS' authorizing legislation mandates that data be made as widely available as practicable (Section 308(c)). (1) However, the mandate to make data available must be guided by NCHS' role as a federal statistical agency and be balanced against the need to protect respondent confidentiality and to assure data quality (U.S. Centers for Disease Control 2002).*

At a global level, the United Nations Statistics Division includes definitions of good practices in their Principles Governing International Statistical Activities (http://unstats.un.org/unsd/methods/stat 1999). Included are the following statements about data dissemination and confidentiality:

> *High quality international statistics, accessible for all, are a fundamental element of global information systems.*
>
> *Individual data collected about natural persons and legal entities, or about small aggregates that are subject to national confidentiality rules, are to be kept strictly confidential and are to be used exclusively for statistical purposes or for purposes mandated by legislation (United Nations 2005).*

How can agencies attempt to address the issue of providing policy makers and researchers with complete and accurate information about the populations for which they make decisions while simultaneously protecting the confidentiality of respondents? In particular, how can statistics play a role? Research and applications concerned with releasing quality information and simultaneously protecting privacy and confidentiality is called statistical disclosure control, statistical disclosure limitation, statistical disclosure protection, and statistical confidentiality, among other names (Fienberg and Willenborg 1998). Statistical disclosure limitation (SDL), or disclosure limitation, is the term we use in this dissertation. It is our goal to develop a disclosure limitation method that provides as much quality information to as many policy makers and researchers as possible while at the same time upholding promises of privacy and protecting confidentiality at an acceptable level.

Two major approaches or disclosure limitation methods generally are implemented to achieve data dissemination with disclosure protection. One method is to limit access to the data. The other is to limit or modify the data themselves (Jabine, 1993). Several methods can be used to implement each approach. These are discussed in Section 1.2.

1.2 Approaches to statistical disclosure limitation

Statistical disclosure limitation is generally approached using one of two broad techniques: limiting access to the data or limiting the data themselves. We review methods to limit access in Section 1.2.1, then review methods to limit the data themselves in Section 1.2.2. This leads us to propose a new method to limit the data, which we describe in Section 1.3.

1.2.1 Limiting access to data

Limiting access to data implies exactly that: access to the data is limited by the agency. The Confidentiality and Data Access Committee (CDAC) of the Federal Committee on Statistical Methodology (FCSM) produces reports on confidentiality and data access issues. The reports summarize methods used in several government statistical agencies and are available on the website at www.fcsm.gov/committees/cdac/resources under "Resources for Confidentiality and Data Access." Included in the collection of reports is one entitled Restricted Access Procedures (FCSM CDAC 2008) that gives an overview of methods currently in use and under development at the U. S. Census Bureau and other agencies.

To access data, generally a user must meet certain demands of the agency before access is granted. If a user does gain access, it is controlled by the agency. Examples include research data centers (RDCs) and secure servers. RDCs are facilities, remote from the agency, where an approved user may access the agency's data. Typically, the agency will allow only trusted data users into secure facilities. In order for a data user to establish *trusted* status, the researcher submits a proposal for work to be done using the agency's data, undergoes a security evaluation, signs a contract, takes an oath, and attends an orientation session. Such a procedure is used by Statistics Canada (Statistics Canada 2008). The trusted user then accesses data at a physically secure computer facility staffed by a Statistics Canada employee, several of which are located throughout Canada. No data or research results may leave the facility without examination and approval. The trusted user promises to produce a research paper upon completion of the proposed research. Similar protocols are standard throughout the U. S. Census Bureau, National Center for Health Statistics, National Bureau of Economic Research, and other agencies. Their facilities exist in various locations throughout the United States.

Research data centers provide researchers with a unique opportunity to access highly detailed, confidential data of high quality. They provide agencies a means to achieve simultaneous data dissemination and privacy protection. Some researchers, however, may find the application process difficult to complete. They also could find that travel to and from and confinement to working in an RDC cumbersome if not expensive. In order to ease the burden of using RDCs, the Cornell Institute for Social and Economic Research maintains a Virtual RDC. The Virtual Research Data Center (VRDC) aims to "assist potential Census RDC users in preparing their proposals, and to train new users in the operating system environment, data and software available on the real Census RDC (VRDC 2008)." The idea is

that researchers can use the VRDC to plan their analyses and work out kinks in preparation for a visit to the actual RDC. At the RDC, the researchers do their final analysis and obtain results from the actual data. This approach minimizes time they must spend at the RDC. Although RDCs and the VRDC provide a means for agencies to disseminate information while protecting the confidentiality of their respondents, the means may not be the optimal solution for all users. At present, data sets are available only froma few surveys.

Remote access servers also limit users' access to the data. Under this scenario, the user submits a query, or requested analysis, to be done using the confidential data. The query is accepted or denied, according to constraints imposed to protect confidentiality. If the query is accepted, the server provides the user with the results of the requested analysis. The user never sees the actual data. Constraints include refusal to requests for a single record, say, and mechanisms to monitor requested analyses so that no combination of them by one user constitutes a disclosure (Gomatam, Karr, Reiter, and Sanil, 2004). Advantages of using a remote access server are that researchers do not need to travel to an RDC, can work remotely, and can conduct many analyses. Also, the researcher can use standard analyses which are based on the original data, and thus have no extra variation or bias introduced from altering, perturbing, or tabulating the microdata (Reiter 2004). On the other hand, combinations of analyses the researcher can obtain results for are limited.

Consider an instance in which a data user submits a request for a table containing frequencies of residents with income over $100,000 in each county of some geographic area and also requests a table containing frequencies of residents with income over $105,000 in the same counties. If the frequency is one less in the $105,000 table than in the $100,000 table for a particular county, then the user knows there is exactly one person with income between $100,000 and $105,000. Adding race as a factor to the tables would allow a user to possibly identify a member of a race category that contains relatively few individuals in a county with a certain income using two tables, one for income over $100,000 by county by race and one for income over $105,000 by county by race. One of these requests may not be a problem with respect to confidentiality protection, but two or more could be problematic.

Another way to limit access to data is to require users to perform secure analyses on distributed data bases. Distributed data bases are data bases that have components held by different individuals or agencies. For example, the U.S. Internal Revenue Service (IRS) might have income and wealth information on individuals, but Census might have survey information from the Current Population Survey. Neither agency might be willing to send

its data to the other agency, but they might under highly controlled circumstances want to do an analysis merging information. This method is used to address the problems specific to agencies sharing the richness of the data without actually sharing the microdata themselves. Secure analyses essentially encode the results after local computations are done and merge encoded results. After the merging, the results are decoded and available to one or more agencies. Data themselves do not go between agencies. Secure analyses on distributed data provide a method for each agency to obtain results using collective data without any actual data sharing between agencies. Reports and papers on this topic can be found at the National Institute of Statistical Science website.

Limiting access to data is an area of statistical disclosure limitation with interesting topics that deserve much attention. This is not within the scope of our research efforts. We investigate a new proposal for limiting or modifying the data themselves. A review of approaches to disclosure limitation by limiting the data themselves follow in Section 1.2.2.

1.2.2 Limiting data

Limiting data is an approach to statistical disclosure limitation in which the data are altered, or perturbed. In this section, we review methods commonly used to limit data. There are several existing methods which have been well studied. Still, there are limitations and there is a desire to provide more informative data.

One approach to limit data is sampling and releasing non-identifying variables. Street address, name, and other detailed information such as birth date or exact age can be identifying variables. Zip code and age range are less likely to be identifying. Sampling the records creates a situation in which a particular record that is released does not necessarily correspond to a single person in the population. Instead, there possibly are many such individuals in the data set, but only those that are sampled are released. Sampling weights can be provided to enable estimation of population quantiles from the sample. The sampling approach has the advantage of releasing microdata; that is, information at the record level. Such information is useful for statistical modeling. In some cases, however, sampling is not viewed as sufficient for protecting confidentiality. As was mentioned, it still is necessary to limit identifying information. Sampling therefore might be used in conjunction with other methods.

A second approach is to restrict the data. Top coding caps the highest value that is recorded. Bottom coding bounds the lowest value that is reported. Top coding could be useful for income or other skew right variables. Bottom coding could be useful for losses or variables that are recorded with negative values, especially if the distribution is skew left. Interval coding replaces a value with an interval code. Interval coding is used often with ages, income, and other continuous variables. It also can be used with discrete variables, especially at the low or high ends of a distribution. Geographic aggregation is the counterpart to interval coding for spatial data. State or county is much less identifying than is zip code or street name and house number. Coding of variables reduces the risk that any individual can be identified by someone using the data set, but it can limit the usefulness of the data set for analysis. The impact on analyses sometimes can be quite severe. Public use versions of several data sets that are made available online to anyone use rather broad coding categories to protect confidentiality.

A third approach is to perturb the data by adding random noise or otherwise distorting the records. For a practically continuous variable such as income or age, one could add a randomly generated error or other perturbation value to each observations recorded value. For discrete or categorical values, one can consider randomly changing values on a set of varaiables from a record in one category to a record in another category. One version of this, referred to as data swapping, swaps values of variables between pairs of records. When ranks are used as the basis of swapping values, the method can be called rank swapping; rank swapping is discussed in more detal in Section 2.3. Swapping can provide confidentiality protection because released values in swapped records do not originally appear in the original record for that respondent. The impact of swapping on both disclosure risk and data utility is discussed in two editions of Lecture Notes in Computer Science (Domingo-Ferrer and Franconi (eds.) 2006) and in a collection of papers in Confidentiality, Disclosure, and Data Access (Doyle, Lane, Theeuwes, and Zayatz 2001). In general, as swapping rates increase, the risk and utility both decrease. When swapping is done between records with similar or proximous values, the method can better preserve statistics while still protecting confidentialiy (Dalenius and Reiss 1978, Schlörer 1981, Reiss 1984).

A fourth approach is to not release microdata but rather to release data in tabular format. Tables are formed by classifying observations as belonging to categories. Categories are based on levels of categorical variables or intervals of continuous variables. Cells of the table contain the frequency of records in each category. They may also contain an aggregate value, or magnitude of a variable of interest. The U. S. Census Bureau, for example,

produces tables of magnitude and frequency. Prior to tabulation, top and bottom coding, rounding, interval coding, noise addition, and swapping are applied to record-level data. The resulting perturbed microdata are then tabulated. Before publishing the resulting tables, threshold rules are used to assess disclosure risk. Rules are designed to protect sensitive cells. Cells are sensitive when they contain a small number of records. Consider a table with a cell with frequency one corresponding to a single unit in the data set. The unit in this cell corresponds to a record with unique values on the variables used to form categories in the table. Unique records are assumed to be easily identifiable, or sensitive, and have high disclosure risk. When sensitive cells are identified, methods including primary and secondary cell supression, controlled rounding, and recategorization are used to alter tables. The resulting table is released with adequate disclosure control. Increasing perturbation before and/or after data are tabulated decreases disclosure risk. The use of these techniques in government statistical agencies is presented in Statistical Policy Working Papers 2 and 22 (Federal Committee on Statistical Methodology 1994).

Categorical data analyses can be used to analyze data in the released tables. Methods include assessing independence using Chi-squared tests and analysis of variance. Researchers can also use logistic regression to analyze effects of variables on the frequency of records in a category. For example, Shlomo and Young (2006) discuss using results from such analyses to measure data utility.

A relatively new approach to limiting data for statistical disclosure control is to create *synthetic* microdata for release. Synthetic data has become a topic in the field of disclosure limitation. Challenges and current research in this field include development of methods to create high quality synthetic data. Measures of disclosure risk and data utility are needed. Current research in synthetic data methods is reviewed in Section 1.3.

1.3 Synthetic data as an approach to disclosure limitation

Synthetic data is an approach to statistical disclosure limitation in which the original data set is replicated, so to speak, with original values replaced by synthetic values. Synthetic values in existing applications typically have been generated using an imputation approach. Imputations have been generated by drawing values from posterior predictive distributions. A posterior predictive distribution is the distribution of the data conditional on the observed data. Intervening between the new, predictive data and the existing, observed data is a sta-

tistical model. One can conceptualize generating synthetic data from a posterior predictive distribution in two steps. First, values of the parameters of a statistical model are drawn from their posterior distribution, i.e., the distribution of the parameters given their prior distribution and the observed data. Second, values of new data are simulated from the data models with values of model parameters equal to the drawn parameter values. The posterior predictive distribution should capture multivariate relationships among variables in the original data set.

In applications to the Longitudinal Employment Household Dynamics program, generalized linear models are used (Abowd and Woodcock 2001, Abowd and Lane 2004). In an application at the U.S. Census Bureau and Duke University, normal linear regression models, binomial or multinomial distributions, and Dirichlet-Multinomial models were used to create synthetic data for the Longitudinal Business Database (Kinney 2007). In an application to group quarters data, a Dirichlet-Multinomial model was used to generate partially synthetic data for the American Community Survey group quarters data (Rodriguez and Hawala 2006). Generalized additive models are also being investigated as a method to generate synthetic data (Hawala 2008). Creating partial synthetic data sets using classification and regression tree models are proposed by Reiter (2005).

The best choice often depends on the specific data that agencies wish to make available in the form of microdata. Often, restrictions must be incorporated to maintain the consistency of the variables. For example, in a data set containing information on household information, the synthetic data should not contain a single person with an age of five years and three dependents. Analytical results based on the synthetic data sets should not differ too much from those based on the original data sets.

The use of multiple imputation ideas with synthetic data generation was suggested by Rubin (1993). By creating multiple synthetic data sets using the same posterior predictive models, one can incorporate and assess the added variability in estimation based on the synthetic data sets. Multiple imputation ideas for use with synthetic data generation are not studied in this dissertation. In principle, the methods proposed here could be implemented multiple times, yielding two or more randomly generated synthetic data sets. Future work could study the use of multiple imputation ideas for variance estimation in the context of our proposed synthetic data method.

In an application at the Iowa Department of Revenue, records with individual income tax

return variables contain values with highly skewed, nonstandard univariate distributions and complicated joint dependencies. Data sets contain continuous and categorical variables. In order to generate a synthetic data set with these properties, we propose combining quantile regression models with hot deck imputation and rank swapping to capture the highly skewed joint conditional distributions. An overview of the proposed method is provided in Sections 1.4 and 1.5. Work done at the Iowa Department of Revenue is introduced in Section 1.6. Further work done at the U.S. Census Bureau is introduced in Section 1.7.

1.4 Quantile regression synthetic data generation

Quantile regression has been used to study economic data by Roger Koenker and others. Quantile regression is analogous to linear regression, which is produced by minimizing squared prediction errors, and median regression, which is produced by minimizing the absolute prediction errors, but it is used to predict quantiles of a response given predictor information. The appeal of this type of regression originates with the nature of variables economists are dealing with. The distributions are often highly skewed, and the relationship between predictors and the mean of a response cannot sufficiently describe their relationship. We initially considered quantile regression for these reasons. The values in our applications that we use to create synthetic distributions have highly skewed, non-standard distributions. It is often plausible that the effect of predictors is not constant throughout the distribution of the response. We use quantile regression models to characterize the relationship of several response variables conditional on several predictors. In fact, we use a series of conditional models in order to retain the joint distribution among multiple response variables. The predictions from the estimated models serve as our synthetic values for each response variable.

1.5 Hot deck imputation and rank swapping as a method to complete data records

We borrow imputation methods from missing data problems to generate values for a set of variables in the synthetic set. Imputation methods applied to missing value problems generally involve one data set containing two sets of records: a set of complete records with values recorded on all variables and a set of incomplete records with values missing on some variables. The goal is to complete the incomplete records using information in all records. Using hot deck imputation in particular, the information in all records is used by comparing

values that are recorded in both the complete and incomplete records and imputing values to an incomplete record from the closest complete record. It is straight-forward to extended this method to generate synthetic data by considering the original data as the set of complete records, and the synthetic data up to the point of quantile regression predictions as the incomplete records.

If we compare values on variables whose values we retained from the original data and variables that we generated using synthetic quantile regression predictions between the original and synthetic data, we can compute distances between synthetic and original records. Then we can impute values from the closest original record into the synthetic record. Following this procedure for every synthetic record, we generate values for this set of variables for every record in the synthetic data set. The result is one synthetic data set with original values on a handful of nonsensitive variables, quantile regression predictions for a set of variables, and original values on remaining variables. By comparing values generated thus far, we hope to retain the joint distribution among the variables we are imputing.

One concern with this method is that imputing values on all remaining variables from a single original record results in too much information from that record being released. It is plausible that a researcher or intruder would use record linkage techniques and either inadvertently or purposefully link the original values with values from the same record from an outside source, thereby identifying the respondent that the particular record belongs to. To lower the likelihood of this type of link, we further perturb the values from the original record using rank swapping.

Rank swapping provides a way to impute values similar to the values of the closest original record, so as not to distort the joint distribution between original, synthetic, and imputed variables. Once the closest original record has been identified, the values on the variables to impute are ranked. Then instead of imputing the value from the closest original record, we randomly select a rank from some distribution centered at the sample rank and impute the value from the original record with the randomly selected rank. The result from doing this, then, is a synthetic set with original values on some nonsensitive variables, quantile regression predictions on another set of variables, and hot deck imputations perturbed by rank swapping on remaining variables.

1.6 Application at the Iowa Department of Revenue

The project motivating this work in disclosure limitation and confidentiality protection originated from a problem faced by the Iowa Department of Revenue (IDR) and Iowa's Legislative Services Agency (LSA). Members of the LSA investigate (presumably, among other things) the effect of proposed tax law changes on the revenue for Iowa, based on individual income taxes. Any proposed tax law change corresponds to a change of values in one or several lines of an income tax return. Given the values on last year's records, say, the revenue for Iowa can be computed by simply altering values in line items corresponding to the proposed change in a tax calculator. If the LSA could manipulate the record-level data and the tax-calculator, its members could compute the state's revenue. However, IDR cannot release individual income tax return data to the LSA because of promises and laws compelling them to protect the privacy of Iowa tax payers. This forces the LSA to submit proposed tax law changes to IDR who then computes the resulting revenue based on past years' data. This occurs several times per year and is incovenient and inefficient for both the IDR and the LSA.

Considering several approaches with IDR, we are pursuing synthetic data as a solution to their situation with the LSA. We combine quantile regression, hot deck imputation, and rank swapping to create synthetic data on several variables. Details of this application are presented in Chapter 4.1.

1.7 Application at the U.S. Census Bureau

The U.S. Census Bureau is among the several government agencies whose purpose is to collect and disseminate data to inform the public and policy makers. Census is bound by strict rules to protect the confidentialitiy of respondent from which data are collected. The American Community Survey (ACS) is administered to collect demmographic and economic data on individuals and households. The database contains rich information on veterans which is of interest to researchers interested in veterans' circumstances. Disseminating this information to researchers in microdata form could provide a useful tool in further research. To address this, we investigated applying the proposed procedure to the ACS data on veterans. Our primary focus was on developing a quantile regression model to accurately predict values for a synthetic data set. In this application, we faced challenges involving restrictions on values for multiple, logically related variables in the same synthetic records. For example, if quantile regression predictions produced a synthetic age value of 17 for a record belonging

to a veteran of World War II, then the age would obviously be inconsistent with the period of military service. Details and results are presented in Chapter 4.2.

1.8 Dissertation outline

The proposed synthetic data method that combines quantile regression with hot deck imputation and rank swapping is presented in Chapter 2. Section 2.1 is devoted to introducing quantile regression, discusses estimation and inference, and presents details on its proposed use to generate synthetic data. Section 2.2 is devoted to hot deck imputation and discusses its typical missing data applications, techniques of applying hot deck, and the proposed use of hot deck in creating synthetic data. Section 2.3 discusses rank swapping for disclosure limitation and its particular use in the proposed synthetic data method. Section 2.4 ties these three methods together to describe the synthetic data method we propose to use to generate synthetic records.

Measurements for disclosure risk and data utility are discussed in Chapter 3. An introduction to disclosure risk is presented in Section 3.1. This includes existing risk measures and notation. The extension to measure risk for a data set generated using the proposed synthetic data method is presented in Section 3.2. We discuss a collection of tools to assess data utility in Section 3.3.

Results from applying our proposed synthetic data method are presented in Chapter 4. An application to individual income tax returns is presented in Section 4.1. Further work done to generate synthetic values for variables in a data set containing records with information on veterans is presented in Section 4.2. Finally, an application of the proposed procedure to a public use microdata set available from the Census Bureau is presented. In this application, we also assess the disclosure risk associated with the resulting synthetic data set. Results are presented in Section 4.3.

Chapter 5 is a conclusion. Some ideas for future work are given.

CHAPTER 2. PROPOSED SYNTHETIC DATA METHODS

Government agencies face demands to release accurate, timely data and to simultaneously uphold their promises of privacy and confidentiality to respondents. We study options for generating synthetic data files for public release. Specifically, we study combining quantile regression and hot deck imputation with rank swapping to produce releasable, usable synthetic microdata. To capture the complex relationships found in demographic and economic data collected by statistical agencies, conditional quantile regression models are used. Predicted values computed from model estimates and key predictors are generated for several confidential variables at random quantiles. Values for other variables are imputed from the original data using hot deck imputation and further perturbed using a rank swapping procedure. The quantile regression predictions are combined with the imputed perturbed values to form a data set with record level data for release that has low disclosure risk and high data utility.

In Section 2.1, we describe the fundamentals of quantile regression estimation and prediction, inference, and our proposed use of quantile regression to generate synthetic data. In Section 2.2, we describe imputation as applied to missing data problems, hot deck imputation techniques, and how it is used in our proposed method to generate synthetic microdata. In Section 2.3, we describe rank swapping as it is generally applied in disclosure limitation and its use in our proposed method. In Section 2.4, we present the combination of methods as the proposed method using quantile regression prediction, hot deck imputation, and rank swapping to generate synthetic microdata for release.

2.1 Quantile regression

Quantile regression has been developed extensively by Roger Koenker and others since 1978 (Koenker and Bassett 1978a, 1978b). Quantile regression research by these authors was inspired by problems in economics, where using the mean to describe and model complex re-

lationships in economics data can be insufficient. Using models for several quantiles can often improve the understanding of such data (Koenker, Fitzenberger, and Machado 2001; Koenker 2000). Research has addressed aspects of estimating and implementing quantile regression and the theory behind it in several frameworks. These include linear (Koenker and Hallock 2001; Koenker and Basset 1982) and nonlinear models (Koenker and Park 1996), smoothing splines (Koenker, Ng, and Portnoy 1994; Koenker and Hendricks 1992), structural equation models (Koenker and Ma 2006), survival analysis (Koenker and Geling 19), and time series models (Koenker and Xiao 2006, 2004; Koenker 2004, 1986; Koenker and Zhao 1996). In several papers and a recent book (Koenker 2005), the authors describe fundamentals and estimation of quantile regression models including inference (Koenker and Xiao 2002; Koenker and Machado 1999), asymptotics (Koenker, Jureckova, Portnoy, and He 1990; Koenker and Bassett 1978), weighted regression and L-Statistics (Koenker and Zhao 1994; Koenker and Portnoy 1989, 1987; Koenker and Bassett 1987, 1982), computation (Koenker 1997; Koenker and D'Orey 1987), and nonparametric approaches. A Vignette with instructions and examples for using the *quantreg* package in R software (Koenker 2005) is also available. Additional contributions to the quantile regression literature, both in collaboration with Koenker and independently, cover topics that include constructing confidence sets for parameters of linear models (Zhou and Portnoy 1996), estimating covariance (Buchinsky 1998), generating prediction intervals (Taylor and Bunn 1999), and modeling counts (Machado and Santos Silva 2003).

The work in this dissertation uses developments presented in the above papers and summarized in Koenker (2005). In the following two sections, we present details from Koenker (2005) to describe fundamentals of quantile regression including estimation and prediction in Section 2.1.1 and inference in 2.1.2. Technical details that are cited are from this book unless otherwise specified. In Section 2.1.3, we propose using quantile regression predictions as a novel method of simulation that can be used to generate synthetic microdata.

2.1.1 Fundamentals of estimation and prediction

Koenker (2005) defines the τ^{th} quantile of Y as $F_Y^{-1}(\tau) = inf\{y : F_Y(y) \geq \tau\}, \tau \in (0,1)$, for random variable Y with right-continuous distribution function $F_Y(y) = P(Y \leq y)$. Koenker and Hallock (2001) provide an alternative definition for sample quantiles from common definitions based on ordering observations. Quantile regression is a method used to fit a curve or surface to a quantile of a response variable at combinations of independent

variables. Similar to least squares regression, in which a curve or surface is fit to the mean of a response variable by minimizing the sum of squared errors, quantile regression fits a curve or surface to a particular quantile of the response variable. In quantile regression, however, the sum of weighted absolute errors, rather than squared errors, is minimized. Weighted absolute errors are defined in terms of a *tilted absolute value function*, denoted $\rho_\tau(u) = u(\tau - I_{[u<0]})$. Consider residual u, where $u = y_i - \hat{y}_i$, then the tilted absolute value function can be thought of as an asymmetrically weighted absolute value for each residual u. Following the notation in Koenker (2005), we write the τ^{th} quantile as ξ_τ, estimated by $\hat{\xi}_\tau$, the solution to $min_{\xi \in \Re} \sum \rho_\tau(y_i - \xi)$.

The above estimation procedure is extended to estimate the τ^{th} conditional quantile function, i.e., to estimate the parameters in a quantile regression model. Suppose the τ^{th} quantile can be expressed as a function of covariates X and parameters β_τ by $\xi(x_i, \beta_\tau) = X\beta_\tau$, where parameters are dependent on the quantile. Then regression parameters, and hence the conditional quantile function, can be estimated by solving $min_{\beta_\tau} \sum \rho_\tau(y_i - \xi(x_i, \beta))$ to obtain $\hat{\beta}_\tau$, and hence $\hat{\xi}(X, \beta_\tau)$.

Technical details of solving the minimization problem $min_{\beta \in \Re^p} \sum \rho_\tau(y_i - \xi(x_i, \beta))$ involve reformulating the sum of weighted absolute errors into a linear programming problem. Artificial variables are introduced to represent positive and negative parts of the residuals, $\{u, v\}$, respectively. The problem becomes one of solving $min\{\tau 1_n^T u + (1-\tau)1_n^T v | 1_n \xi + u - v = y\}$, where 1_n is a vector of ones. The process of converting the quantile regression estimation into a linear programming problem has been automated in the contributed R package *quantreg,* authored by Koenker and described in detail in the Vignette (Koenker 2005). The author describes four automated algorithms that can be used to solve this minimization and produce quantile regression estimates. They are a modified version of the Barrodale and Roberts simplex algorithm for $l1$-regression, the Frisch-Newton algorithm, the Frisch-Newton algorithm with a preprocessing step, and the Fitzenberger implementation of Powell's censored quantile regression estimator. A user can also specificy linear inequality constraints on the fitted coefficients. Details on implementation can be found in the Vignette, Koenker (2002, 2005), and Koenker and Hallock (2001).

With knowledge of the fundamental estimation procedures in hand, we now describe its typical use. Suppose a data set contains records with observed values on variables X and Y. Consider modeling the τ^{th} quantile of Y given values of X using the expression $Y_\tau = \xi(X, \beta_\tau) + \epsilon_\tau$. Here, $\xi(X, \beta_\tau) = X\beta_\tau$ is a linear function of X, parameters β_τ, and ϵ_τ,

independent random errors. Using the methods described above, we can obtain parameter estimates, $\hat{\beta}_\tau$, for some quantile $\tau \in (0,1)$. Further, using $\hat{\beta}_\tau$ and X values, predicted values of Y_τ, $\hat{y}_\tau$, can be computed where $\hat{y}_\tau = X\hat{\beta}_\tau$. If one had several responses y at each value x, then one could evaluate the predictions of quantiles. For example, suppose that X takes on many values, but at value x_0 there are $1,000$ response values y. If one is interested in, for example, the 90^{th} quantile, then the *quantreg* function can be used to estimate $\beta_{0.90}$. The prediction at x_0 of the 90^{th} quantile is $x_0\hat{\beta}_{0.90}$. If the model is accurately estimating the 90^{th} quantile conditional on X at this value of x, then there should be about 100 response values y above the predicted value and 900 below.

Statistical inference for regression quantile estimation are presented in Section 2.1.2. Inference can be performed for the estimated coefficients and for the predicted quantiles themselves. In Section 2.1.3, we give details of our proposed disclosure limitation method that incorporates quantile regression as a tool to create synthetic microdata.

2.1.2 Inference

A brief summary of results concerning statistical inference in quantile regression models is presented here. Details and further results can be found in Koenker (2005). Only those relevant to our application are presented in this section. If $Y_1, ..., Y_n$ are identically and independently distributed random variables with cumulative distribution function F, assuming F has continuous density f at ξ_τ and that $f(\xi_\tau) > 0$, then for some quantile $\xi_\tau = F^{-1}(\tau)$, the objective function of the τ^{th} sample quantile is the sum of convex functions and is itself convex. This implies that its gradient, g_n is monotone increasing in ξ, i.e. $\hat{\xi}_\tau > \xi$ if and only if $g_n(\xi) < 0$, and

$$P\{\sqrt{n}(\hat{\xi}_\tau - \xi_\tau)\} = P\{g_n(\xi_\tau + \delta/\sqrt{n})\} = P\{n^{-1}\sum(I_{[Y_i<\xi_\tau+\delta/\sqrt{n}]} - \tau) < 0\}.$$

As Koenker (2005) describes, this implies that the asymptotic behavior of $\hat{\xi}_\tau$ can be reduced to a DeMoivre-Laplace central limit theorem problem, which leads to the following convergence result:

$$\sqrt{n}(\hat{\xi}_\tau - \xi_\tau) \rightharpoonup N(0, \omega^2) \text{ as } n \to \infty$$

where $\omega^2 = \tau(1-\tau)/f^2(\xi_\tau)$. This result means that the estimated quantile in large samples has a distribution that is approximately normal. The mean is the actual quantile. The variance is related to a Bernoulli variance ($\tau(1-\tau)$) divided by the squared density of Y at

the quantile, which is analagous to a sample size.

This result can be extended to the approximate asymptotic joint distribution of several estimated quantiles. Setting $\varsigma_n = (\xi_{\tau_1}, ..., \xi_{\tau_m})$ with estimates $\hat{\varsigma}_n = (\hat{\xi}_{\tau_1}, ..., \hat{\xi}_{\tau_m})$, then

$$\sqrt{n}(\hat{\varsigma} - \varsigma) \rightsquigarrow N(0, \Omega) \text{ as } n \to \infty$$

where $\Omega_{m \times m}$ is a matrix with the elements

$$(\omega_{ij}) = \frac{\tau_i \wedge \tau_j - \tau_i \tau_j}{f(F^{-1}(\tau_i)) f(F^{-1}(\tau_j))}.$$

The expression $\tau_i \wedge \tau_j$ is evaluated to equal τ_i when $i \leq j$ and τ_j when $i > j$ (Serfling 1980).

In particular, we are interested in distributional properties of predictions based on models of the form $Y_\tau = X\beta_\tau + \epsilon_\tau$. If the $\epsilon_{\tau,i}$ have common distribution function F with associated density f, such that $f(F^{-1}(\tau_i)) > 0$ for $i = 1, ..., m$, and $Q_n = n^{-1} \sum x_i x_i^T$ converges to Q_0, a positive definite matrix, then the m p-variate quantile regression estimates $\hat{\varsigma}_n = (\hat{\beta}_n(\tau_1), ..., \hat{\beta}_n(\tau_m))$ are asymptotically Normal,

$$\sqrt{n}(\hat{\varsigma}_n - \varsigma_n) \quad = \sqrt{n}(\hat{\beta}_n(\tau_j) - \beta(\tau_j))_{j=1}^m \quad \rightsquigarrow N(0, \Omega \otimes Q_0^{-1}).$$

Provided the assumptions above are reasonable, one can extend the convergence properties of $\hat{\varsigma}_n$ to quantile regression predictions $\hat{Y}_\tau = X\beta(\tau)$. When $\tau = (\tau_1, ...\tau_m)$, $m \leq n$, then

$$\sqrt{n}(X\hat{\beta}_\tau - X\beta_\tau) \rightsquigarrow N(0, X^T \Omega \otimes Q_0^{-1} X) \text{ as } n \to \infty.$$

Koenker (2002) explains that if the design matrix X is assumed to satisfy Conditions D1 and D2, then Condition F is a necessary and sufficient condition for the consistency of quantile estimates $\hat{\beta}(\tau)$. The conditions are restated here.

Condition F The τ^{th} conditional quantile function of Y given X can be written as $X\beta(\tau)$ and conditional distribution functions of Y_i, F_{ni}, satisfy $\sqrt{n}(a_n(\epsilon) - \tau) \to \infty$ and $\sqrt{n}(\tau - b_n(\epsilon)) \to \infty$, with $a_n(\epsilon) = n^{-1} \sum F_{ni}(x_i^T \beta(\tau) - \epsilon)$ and $b_n(\epsilon) = n^{-1} \sum F_{ni}(x_i^T \beta(\tau) + \epsilon)$.

Condition D1 $\exists d > 0$ such that $lim\ inf\ inf_{||u||=1}\{n^{-1} \sum I_{[|x_i^T u| < d]}\} = 0$

Condition D2 $\exists D > 0$ such that $lim\ sup\ sup_{||u||=1}\{n^{-1} \sum (x_i^T u)^2\} \leq D$

Condition D1 is needed for identifiability. Condition D2 controls the rate of growth of $\{x_i\}$, which is satisfied under the condition that $n^{-1}\sum x_i x_i^T$ converges to a positive definite matrix. If these assumptions are too stringent, Conditions D1 and D2 can be relaxed, Condition F can be strengthened, and consistency of $\hat{\beta}(\tau)$ will still hold. For details, see Koenker (2002, 2005) and Zhao, Rao, and Chen (1993).

The Conditions D1, D2, and F generally hold for large n. We use these inference results to develop a measure of disclosure risk in a synthetic data set created using quantile regression predictions. This is discussed in Chapter 3.

2.1.3 Quantile regression for synthetic data

Here we propose using quantile regression predictions to generate synthetic values for statistical disclosure limitation. Suppose we wish to release data set with variables X and Y for public use, but must protect confidentiality. To do so, we wish to generate a synthetic data set for release with properties very similar to the original data. Suppose variable X is non-sensitive and original values can be released without concern, but that variables $Y_1, Y_2, ..., Y_s$ are sensitive and must be protected. We propose using quantile regression to generate synthetic values for $Y_1, Y_2, ..., Y_s$ such that the joint distribution of $Y_1, Y_2, ..., Y_s|X$ in the original data is preserved.

If we write the joint distribution of $Y_1, Y_2, ..., Y_s$ given X as $Y_1, Y_2, ..., Y_s|X$, we can rewrite it using a sequence of conditional distributions as follows:

$$Y_1, Y_2, ..., Y_s|X = (Y_1|X)(Y_2|X, Y_1)\cdots(Y_s|X, Y_1, Y_2, ..., Y_{s-1}).$$

That is, given distributional assumptions about $Y_1, Y_2, ..., Y_s|X$ and values for X, we could generate values for Y_1 conditional on X, values for Y_2 conditional on X and Y_1, and so on. For each variable Y_i, we propose to use quantile regression to model the relationship between the quantiles of variable Y_i and variables X and Y_j $(j < i)$. Once the coefficients are estimated, for each observation for variable Y_i, we will generate a random quantile (a value between 0 and 1) and the predicted quantile value conditional on values of X and Y_j $(j < i)$.

The quantile regression predictions at several quantiles should be able to accurately represent the original data in the sense that the conditional distributions should be roughly equivalent. The predictions also should be releasable without compromising confidentiality

because, except for nonsensitive X variables, no actual values are being released.

The method involves fitting quantile regression models for each variable using conditional models $Y_1|X$, $Y_2|X, Y_1$,..., $Y_s|X, Y_1, Y_2, ..., Y_{s-1}$, and computing predicted values from the parameter estimates, values on X, and newly generated synthetic values on variables generated ealier in the sequence. Specifically, we generate synthetic values for $Y_1|X$ first, denoted $Z_1 = \hat{y}_{1,\tau} = X\hat{\beta}_{1,\tau}$. Then, using estimates $\hat{\beta}_{2,\tau}$, obtained from fitting $Y_2 = (X,\ Y_1)^T\beta_{2,\tau} + \epsilon_\tau$, we generate synthetic values $Z_2 = \hat{y}_{2,\tau} = (X,\ Z_1)^T\hat{\beta}_{2,\tau}$, and so on. We fit the models for each Y_l based on values in the data so that our parameter estimates, $\hat{\beta}_{l,\tau}$, represent the relationships between Y_l and predictors $X, Y_1, Y_2, ..., Y_{l-1}$ that appear in the original data accurately. The synthetic predictions are computed using the parameter estimates and the sythetic values in order to preserve those relationships in the synthetic data.

A simplification could occur if not all variables are needed in later predictions. In general, such simplifications correspond to assumptions of conditional independence between variables. For example, if Y_1 and Y_3 are conditionally independent given X and Y_2, then $Y_3|X, Y_2$ is a sufficient model to predict Y_3. Practically, we may make similar simplifications when parameter estimation is too computationally intensive for large models. Note that the sequential procedure is designed to be expedient. An alternative would be to build a full model for the joint distribution of all variables and simultaneously generate a vector of values. In most large data sets this will be prohibitively difficult. Future extensions could examine intermediate options between the sequential procedure adopted here and something closer to sampling from the full joint distribution.

Up to this point nothing has been said about which quantiles are used to compute synthetic values. We can consider this from two perspectives: accuracy and disclosure control. On one hand, our goal is to generate synthetic data that accurately represents the original data. From this perspective, we might consider estimating the quantile of each observation with respect to the other values on that variable or we could even consider modeling the quantiles for each variable with respect to other variables in the data set. From the perspective of disclosure control, we are actually aiming to introduce randomness into the records so that no one record will be identifiable. Here, we propose randomly selecting quantiles at which to generate values. In this dissertation, we generate synthetic values at a randomly selected quantile for each variable on each record. In future work, it would be interesting to study other approaches to quantile selection, including those listed above.

The proposal to generate random quantiles for imputation or prediction of variables Y_1 through Y_s does not mean that nonsensical records will be created. If Y_1 was generated based on a quantile regression involving X as a predictor and Y_2 was generated based on a quantile regression involving only X as well, and if Y_1 and Y_2 were not conditionally independent, then nonsensical relationships would reults. If Y_1 and Y_2 are highly correlated, then a high quantile of Y_1 should be associated with a high quantile on Y_2 (for a given values of X). Our method should achieve this same relationship. If the quantile for Y_1 given X is high, then the values of Y_1 will be large relative to other cases with the same value of X. Then given X and Y_1, in general due to positive correlation a high value of Y_2 should be predicted. Within the population of untis with the given value of X and a relatively high value of Y_1, the value of Y_2 should be able to vary within the range of the conditional distribution. On average, the value of Y_2 will be at the median of the conditional distribution given X and Y_1. The main point is that our method should preserve correlations and relationships as long as the proper conditioning is included in the quantile regression models.

To summarize, we propose generating synthetic values for each record using conditional quantile regression models at a randomly selected quantile on each record for each variable. We denote the original data on non-sensitive variables X and the original data on sensitive variables Y. Values on variable X are released as they appear in the original data set. Synthetic values Z are generated for each record on variables in Y. For each variable $Y_1, ..., Y_s$ in Y, we randomly select quantiles for each record and denote them as $\tau_1^*, ..., \tau_s^*$, where $\tau_l^* = (\tau_{l1}^*, ..., \tau_{ln}^*)$, for variable l, $l = 1, ..., s$, indices for records 1 through n. Using quantile regression models for each value in τ_l^*, we estimate parameters in the model(s) $Y_l = (X\ Y_1\ \cdots\ Y_{l-1})^T\beta(\tau_l) + \epsilon_{\tau_l}$. Using estimates from the fitted models, we compute predicted, synthetic values Z_l, where $Z_l = \hat{y}_{l,\tau_l^*} = (\hat{y}_{l1,\tau_{l1}^*} \cdots \hat{y}_{ln,\tau_{ln}^*})$.

2.2 Hot deck imputation

Imputation methods are typically implemented to deal with missing value problems. Consider a data set that contains several records with information on a number of variables. Some of the records are complete, with values recorded for every variable. Some records are incomplete, with values recorded for some variables but missing for others. Imputation is a method used to fill in the missing values of incomplete records based on all of the recorded values in the data set. In Section 2.2.1, we discuss imputation in general, as applied to missing value problems. In Section 2.2.2, we discuss hot deck imputation in particular. In

Section 2.2.3, we describe the proposed use of hot deck imputation for statistical disclosure limitation (SDL).

2.2.1 Imputation for missing value problems

Little and Rubin (2002) give an overview of missing data patterns and techniques used to address their presence. Missing values might occur according to some *mechanism* or *process* with respect to variables in the data set. Depending on what the mechanism is, appropriate estimation and inference should be used. As we present our proposal to implement imputation for SDL in Section 2.2.2, it becomes clear that determining a missing data mechanism is a topic we are not concerned with. Instead we will be concerned with appropriate analyses on the synthetic data set and will address this in Sections 2.2.2 and later in Section 3.3 where we discuss data utility.

Imputation procedures provide a means by which to fill in missing values in incomplete records based on information from complete records. Little and Rubin (2002) introduce imputed values as "means or draws from a predictive distribution." Computing means and/or drawing values requires some knowledge and assumptions about the predictive distribution of the missing values. The authors suggest *explicit modeling* and *implicit modeling* as two approaches to generating values from a predictive distribution. Using explicit modeling, a predictive distribution is based on a well-known and understood statistical distribution. Under explicit modeling the missing data might be imputed via mean, regression, or stochastic regression imputation. Using implicit modeling, values are generated from a predictive distribution implied by some algorithm. Under implicit imutation, missing data might be imputed via hot deck imputation, substitution, or cold deck imputation. In some situations, a combination of implicit and explicit modeling might serve to impute in the best way.

To select an appropriate imputation procedure, we consider implementing the procedure, the information in the original data, and the goal we wish to reach using imputation to fill in missing values. If we choose to perform regression imputation, our imputed values would be predicted values obtained from a fitted regression model. Using stochastic regression imputation, we would again compute predicted values, but also add residuals to predictions to obtain imputed values. Either of these methods would be reasonable and generate sound imputations, provided the model used accurately reflects the true model the data are generated from. Under implicit modeling, we fill in values according to some algorithm.

Using hot deck imputation, values from complete records are imputed to fill in missing values on incomplete records. Substitution is a method that adds additional respondents to the data set to substitute these complete records for the original incomplete records. Cold deck imputation imputes a value from a reference table for each variable with missing values.

Consider our proposed synthetic data method, up to using quantile regression as described in Section 2.1.3. We propose to keep values on some variables and to compute predicted values for a set of sensitive variables. Our predicted values come from quantile regression models. If we consider our original data as the set of complete records and our synthetic data as our set of incomplete records, then we can fit the proposed method from Section 2.1.3 into a missing data framework. Namely, we could describe quantile regression predictions as regression imputations for the "missing" values on sensitive variables. After this point, we might consider using an alternate imputation procedure to fill in "missing values" on the remaining variables.

We chose to generate synthetic values for a set of sensitive variables using quantile regression based on our belief that doing so would accurately reproduce the conditional joint distribution of the sensitive and non-senitive variables. The quantile regression method is computationally intensive, so we turn to hot deck imputation, a less demanding procedure to generate values for remaining variables that will preserve the conditional joint distribution of variables, though likely not as well as the quantile regression predictions. Our reasoning and a detailed description are presented in Section 2.2.1. In Section 2.2.2, we describe the proposed use of hot deck imputation for SDL.

2.2.2 Hot deck techniques

In this section, we consider various options for using hot deck imputation based on Little and Rubin (2002). Recall the missing data problem described in the previous section: our data set contains some complete records with values recorded for all variables, and some incomplete records, with missing values for some variables and observed values on other variables.

Hot deck imputation relies on matching records based on observed values on all records in the following way. In Little and Rubin (2002), complete records are referred to as donors and incomplete records are referred to as candidates. Each candidate is compared to all

possible donors. The intention is to impute values from the donor that is the closest to the candidate. There are several approaches to measure closeness between the donors and each candidate. Closeness, or inversely, distance between records can be measured with respect to values on either a single variable or several variables. In some highly restrictive situations we might consider imputing values to a candidate only from donors that exactly match the candidate on some set of variable values. We could also loosen this constraint to allow imputing values from donors that match the candidate within some range, or distance of the candidate. We consider close matches as records with relatively small distances to the candidate. In this sense, we would identify a donor as an exact match to the candidate if the distance between them was zero. In situations where exact matching is not an option or where it is desirable to have more than one options of donor record, we compute positive valued distances between the candidate and the donors. Measures of distance between records and selecting the donor to impute from are described in the remainder of this section.

In order to compute the distance between donors and candidates, we require a distance metric. To choose among several possible donors, we require a selection process. We discuss possible distance measures and then a selection procedure based on the measures. In general, we denote the distance between record i and record j as $d(i, j)$. The closest matching donor record to the candidate would have the smallest $d(i, j)$ value. An exact match would have $d(i, j) = 0$. There are several measures used in hot deck imputation to identify nearest neighbors, i.e. to define $d(i, j)$.

Three possibilities are given in Little and Rubin (2002) for *nearest neighbor matching*. The options define distance based on either the predictive mean of y, $\hat{y}(x)$, or directly on values on the observed variables x. The variable y is assumed to be univariate, but the predictor variable x could be a vector. The predicted values are produced by a model with its parameters estimated. Estimation of parameters is accomplished using the records that are complete on variables y and x. Predictions can be made then for all records with x observed. The three distance measures are as follows:

Predictive mean matching	$d(i,j) = [\hat{y}(x_i) - \hat{y}(x_j)]^2$
Maximum deviation	$d(i,j) = max_k \lvert x_{ik} - x_{jk} \rvert$
Mahalanobis distance	$d(i,j) = \sqrt{(x_i - x_j)^T S_{xx}^{-1} (x_i - x_j)},\ S_{xx} = cov(X).$

Predictive mean distance(and matching based on it) accommodates a vector of predictors x that contains different types of variables, such as continuous and categorical. Predictor

variables x_k that do not predict y well or do not make much difference in predictions then have little impact on matching, whereas variables that impact predictions a lot matter the most. The x variables can be on very different scales as well; the magnitudes of coefficients adjust based on how x relates to y. Maximum deviation distance makes the most sense to use when the x variables are on similar scales and are quantitative. Mahalanobis distance also makes the most sense with quantitative variables. The use of the covariance matrix means that the x variables can be measured on different scales and can be correlated. Of course, there are other distance measures that could be employed.

Of these metrics, we prefer Mahalanobis distance because of its generally desirable properties for measuring statistical distance and because it can accommodate one or more variables. As opposed to predictive mean distance, one does not have to compute predictions. In our application, we anticipate taking values on several variables from one donor. As a result, we do not anticipate wanting to match on only one outcome variable y. In Section 2.2.3, we describe our proposed use of hot deck imputation for statistical disclosure limitation.

Two further comments can be made concerning hot deck imputation. First, if several potential donor records have the same distance to the candidate, as could occur with discrete matching variables x, then one can randomly select a donor from the set of closest potential donors. Second, if one has a mix of discrete and continuous matching variables, one could require exact matching on the descrete variables and then compute a distance based on the continuous variables. This allows exact matching on variables such as gender and state of residence, but distance-based matching on variables such as age and income.

2.2.3 Hot deck imputation for synthetic data

Consider again our proposed synthetic data method, up to using quantile regression as described in Section 2.1.3. As we discussed in Section 2.1, after implementing the quantile regression procedure, our synthetic data set contains original values on the handful of nonsensitive variables and synthetic values for variables that we have generated quantile regression predictions for. Now, if we vertically merge the original data and the synthetic data, the result is one data set with several complete records (original data) and several incomplete records (synthetic set so far). The number of rows of the resulting data set is twice the original. Once we c omplete the data in the synthetic rows, then we will discard the orignal data in the top half of the data set. If we consider the remaining variables to

have missing values in the synthetic, or incomplete records, then we can use the hot deck procedures described in Sections 2.2.1 and 2.2.2 to impute values from the original, complete records into the synthetic records.

In our synthetic data method, we compute the Mahalanobis distance between each synthetic record and all original records, with respect to the original nonsensitive variables and the variables for which quantile regression predictions were generated. For each synthetic record j, we compute $d(i,j) = \sqrt{(x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)}$, where $S_{xx} = cov(X)$, for every original record i.

For an application in which we retain original values on categorical nonsensitive variables, we require exact matching between the synthetic and original values and compute Mahalanobis distance where the x_i represent quantile regression predictions on synthetic record i and x_j represent values on original record j in the formula above. If the categorical nonsensitive variables have many levels and exact matching on them provides a very small number of records from which to impute, we might have concern that the record selected for imputation will be too much like the original record and hence vulnerable to identification. In this case, we can form broader categories by combining levels of the nonsensitive variables and then use exact matching within the recategorized nonsensitive variables. This should provide a larger number of records from which to impute, decreasing the vulnerability of each record. Alternatively, if we retain values on continuous nonsensitive variables, we could form intervals and do exact matching based on the interval that the nonsensitive values lie in, and then compute the Mahalanobis distance between quantile regression predictions and original values on records with nonsensitive values in the same interval.

With the distances calcluated for record j in the synthetic set, we can imagine imputing values from record k in the original set if $d(k,j)$ is the smallest compared to all other $d(i,j)$, $i = 1, ..., n$. If more than one record has distance equivalent to the smallest, some selection procedure is necessary to choose a record to impute from. Note that we do not actually propose imputing all values from a single original record. Details are presented in Section 2.3.

Various selection procedures could be implemented. For synthetic records with several equidistant original records, additional variables could be used to measure distance. Random selection from among the equidistant originals could be employed. In our procedure, we propose to select an original record randomly from among several closest records. This is a different procedure than is usually used for missing values implementation. In missing value

problems, having imiputations as close as possible to the original has advantages. Of course, in a missing data situation, ultimately we want to produce valid statistical inferences for population quantities of interest. In statistical disclosure limitation, the added randomness from random selection from among close donors could provide added confidentiality protection.

To summarize, we compare original and synthetic records with respect to values on variables that we generated quantile regression predictions for. From among the closest m original records, say, we randomly select record k. We refer to record k as the identified *match* for synthetic record j. If we imputed values for all remaining variables from matching record k, the result would be complete record j. However, since our interest is not only in preserving the joint conditional distribution of sensitive variables, but also in protecting the confidentiality of respondents, or records, including too many original values from a single record could result in an increased disclosure risk. With several variables from a single record, record linkage techniques could be used to link the information on these imputed variables with outside information. An intruder might perform such linkage and declare identification of a respondent. Even if synthetic values are not equal to the "identified" respondent's values on these variables, the values on the imputed would be [nearly] equal to the respondent's and an announced "identification" could be just as harmful to both the respondent and the agency. As we discuss in Chapter 3 of this dissertation, an ultimate goal of any proposed statistical diclosure limitation approach is to dissuade any identification, whether true or false.

Our approach to minimize the likelihood of identifying a respondent is to perform rank swapping on the original record identified as the hot deck match. Rank swapping is discussed in detail in Section 2.3. Once all values are imputed for the synthetic data set, we discard the original data and are left with a synthetic data set containing n records.

2.3 Rank swapping

2.3.1 Rank swapping for disclosure limitation

Rank swapping and swapping in general are methods attributed to Dalenius and Reiss (1978, 1982). They proposed swapping as a technique to be applied to original data records for disclosure limitation purposes. Suppose a data set has information on sex of respondent,

race, county, and income. In some counties, there might be few individuals with a certain sex.race combination. Releasing a data set with all four variables unaltered could essentially announce the income of certain individuals. Swapping sex of some respondents could provide some protection by putting some doubt as to the authenticity of a sexrace combination. Generally, a certain percent of the data set is randomly selected and values on a variable are randomly swapped or excchanged between units. The percent is often not too large, because a large percentage swap would distort the data set too much for inferential purposes. On the other hand, a percentage that is too small will not provide much protection. Swapping values on records could also be cone for pairs of variables, such as sex and race simultaneously. Doing so would provide even more protection for individuals with rather uniqe sex-race values in some counties.

Rank swapping is mostly useful for quantitative variables. For a variable like income, if two incomes are swapped, then a very large income could be replaced with a very low income, and the inferential usefulness of the data set could be comopromised. Also, the apparent inconsistency between a small swapped income and having considerable financial assets or living in a wealthy area could undo much of the disclosure protection. Instead, swapping income with individuals whose income is of a similar rank to the original could provide some disclosure protection without creating so much distortion. Rank swapping then is the process of randomly swapping values between units with values of similar rank.

The authors consider frequency tables formed using a data base with individual records on categorical variables, and aim to swap values between records while preserving *t-order* statistics from tables formed using the original data, where t is the number of variables used to compute marginal frequencies. Theorems and proofs are presented to argue that swapping algorithms that preserve t-order statistics produce data for release that are protected under various assumptions about the expected number of swaps and the size of the database. Schlörer (1981) defines Dalenius and Reiss (1978) rank swapping as a multidimensional transformation applied to the original database. Schlörer formulates structural requirements that permit such transformations. Reiss (1984) expands on original swapping ideas to consider *approximate data swapping,* a method used to produce swapped data that *approximately* preserves t-order statistics, allowing for less stringent swapping algorithms. He provides algorithms to perform approximate data swapping and shows that this method protects data to a specified degree of confidentiality through a simulation study.

In a Census Bureau Report, Moore (1996) extends initial results in Dalenius and Reiss

(1978) to propose rank-based proximity swapping as an approach to disclosure limitation. Rank swapping here is designed to satisfy one of the following two constraints:

1. maintain the correlation between two variables, one of which is swapped, to be within some factor of the original correlation; and

2. maintain each value in a swapped record to be within some specified distance of its original value.

The author also presents formulas used to compute swapping parameters and to assess computation time, supported by theoretical justification and results from an application to the 1993 Annual Housing Survey (AHS) Public Use File. Data swapping was actually used at the Census Bureau in the 2000 Census to swap records before creating tables for release. Details on this can be found in Zayatz (2005, 2007). Working Paper 22 (2005) emphasizes that the purpose of implementing swapping is to introduce uncertainty so that users or intruders do not know whether the values in each released record correspond to a particular target with certainty.

In Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies (2001), Domingo-Ferrer and Torra compare disclosure limitation techniques and find that rank swapping performs quite well in comparison to several other methods, when the best swapping parameter is chosen. Working Paper 22 (1994, 2005) also includes swapping and rank swapping as statistical disclosure limitation techniques used by agencies. In the revised (2005) edition, the authors add that swapping can produce data sets in which some statistics can be preserved by placing restrictions on swapping parameters. The report also discusses a web based software package available through NISS, through which a user can upload their data file which gets swapped through distributed computing as part of the NISS Web Services, and then download a swapped file. Additional details can be found in Sanil, et al. (2003) and on the NISS website. Along the same lines of restricted swapping, Takemura (2002) describes local recoding and record swapping within pairs of records deemed close. Closeness is evaluated by computing distance between records as a function of weighted absolute differences. The author also suggests that swapping within categories, rather than pairs, could improve disclosure limitation but notes limitations due to optimization associated with minimizing the absolute differences.

A collection of proceedings from the Computational Aspects of Statistical Confidentiality (CASC) 2004 conference are published in Lecture Notes in Computer Science (LNCS)

in a volume entitled Privacy in Statistical Databases. In one article, Feinberg and McIntyre review Dalenius and Reiss' paper introducing swapping and extensions by Dalenius and Reiss, as well as other variations in the literature. A second edition of Privacy in Statistical Databases in LNCS was published in 2006. It contains proceedings from the Center of Excellence in Statistical Disclosure Control (CENEX-SDC) conference. In this edition, Trottini, et al. (2006) present *microaggregated swapping*, a procedure used to swap values on sensitive variables. Microaggregated swapping permutates the original data and corresponding sampling weights to produce releaseable data. They apply this method to the Household Expenditure Survey, collected by the Italian National Statistical Institute. Authors Muralidhar, et al. (2006) present swapping based on proximity of the ranks of confidential variables. Using their procedure, each value with rank i on a confidential variable is swapped with the value in another record, with rank j, where i and j are within some predetermind value, or *swapping distance*. The authors also present *data shuffling* as an alternative to swapping, where instead of swapping values of ranks i and j for release, values are shuffled, so that the value with rank i is substituted for the value with rank j whose value is substituted for the value with rank k and so on.

In a recent paper by Nin, Herranz, and Torra (2008), the authors argue that standard rank swapping procedures used for disclosure limitation actually produce data sets that are subject to greater disclosure risk than previously believed. They develop a record linkage procedure specifically designed to link records that have been swapped, resulting in more accurate matching on rank-swapped data than standard linkage procedures had previously produced. In response to this, they develop two updated rank-swapping procedures that are more immune to both standard record linkage methodology and their methodology specifically designed to identify records that had been swapped.

In this literature, swapping is implemented as a primary disclosure limitation method. In our proposed method, described in detail in Section 2.3, rank swapping supplements hot deck imputation to generate values for the synthetic data set. Future research could address other possible swapping techniques and compare our proposed methods to alternatives in terms of disclosure risk and data utility.

2.3.2 Rank swapping for synthetic data

Consider our proposed procedure up to this point using quantile regression and hot deck imputation. At the end of Section 2.2.3, we noted that imputing values on all variables remaining after producing quantile regression prediction could have the undesirable effect of increasing the likelihood of a record being linked with some outside information. In order to introduce randomness into the synthetic data on the variables for which we impute values, we propose implementing a rank swapping procedure.

Suppose for a single synthetic record, we have identified the original record to impute values from using the hot deck procedure. Instead of imputing values on all remaining variables from the single identified record, we compute the rank r of each variable with respect to values in the original data set. We swap rank r with randomly selected rank r^* and impute the value corresponding to rank r^* in the original records. The rank r^* is selected independently for each remaining variable.

We use a proximity swap so that imputed and predicted values have similar joint distributions as the original data. We could specify the proximity in several ways, by fixing the distance from the original rank, limiting swapping between pairs or larger groups, limiting swapping within categories, or drawing the swapped rank from within some interval. We limit swapping to occur within categories defined by the non-sensitive variable values and, for the sake of introducing another level of randomness to our synthetic data, we randomly select the swapped rank from an interval centered at the original rank with equal probability, i.e. $r^* \sim \text{Uniform}(r - \delta, r + \delta)$. Other distributions and various values for δ can be used, depending the desired proximity of the swapped rank. Future work might investigate the effects that different swapping methods and parameters, such as distribution and proximity from which ranks are drawn, have on the data utility and disclosure risk of the resulting synthetic data set.

2.4 Proposed synthetic data method: a combination of techniques

In Sections 2.1, 2.2, and 2.3, techniques for creating synthetic data using quantile regression predictions, hotdeck imputation, and rank swapping were presented. In this section, we combine methods to present the entire proposed synthetic data method. The details

for each portion or step of the proposed procedure are presented in the preceding sections. Here we tie them together to describe a clear picture of our proposed method. The notation used below parallels Reiter (2005) used to formulate disclosure risk measures. Details on disclosusre risk are presented in Chapter 3 of this dissertation.

Suppose we denote our original data set as Y, which contains variables $Y_0, Y_1, ..., Y_s, X$. Variables in X are non-sensitive and their original values are released. Variables Y_0 are unique identifiers and are never released in any form. Variables $Y_1, ..., Y_s$ are sensitive. We wish to create a synthetic data set Z for release including values for variables $Z_1, ..., Z_s$, and X on every record. Variable Z_i is the synthetic (artificially generated or imputed) version of variable Y_i. Suppose we generate values for variables $Z_1, ..., Z_f$ using the quantile regression procedure and values for $Z_{f+1}, ..., Z_s$ using hot deck imputation with rank swapping. We describe the details below.

In Figure 2.1, we present an illustration of the original and synthetic data sets. In this illustration, variables containing information are shaded, and variables for which we will generate values as clear, or white. Figure 1 shows the first step in our procedure, where we essentially copy original values on non-sensitive variables X into the synthetic data set using no disclosure limitation procedure (no SDL). Up to this point, the synthetic set contains only values for X. In the next step, we generate values for $Z_1, ..., Z_f$ using quantile regression predictions (QR).

To generate values for $Z_1, ..., Z_f$, we use the sequential conditional models discussed in Section 2.1.3. We begin by generating values for Z_1. We use X and Y_1 from the original data to fit quantile regression models $Y_{1,\tau_1} = X\beta_{1,\tau_1} + \epsilon_{\tau_1}$ for randomly selected quantiles τ_1^* for each record. Using estimates $\hat{\beta}_{1,\tau_1^*}$, we compute synthetic values $Z_1 = \hat{y}_{1,\tau_1^*} = X\hat{\beta}_{1,\tau_1^*}$. Similarly, we generate values for Z_2, using X, Y_1, and Y_2 from the original data set to fit quantile regression models $Y_{2,\tau_2} = (X,\ Y_{1,\tau_1})^T\beta_{2,\tau_2} + \epsilon_{\tau_2}$ using randomly selected quantiles τ_2^* for each record. Using estimates $\hat{\beta}_{2,\tau_2^*}$, we compute synthetic values $Z_2 = \hat{y}_{2,\tau_2^*} = (X,\ Z_1)^T\hat{\beta}_{1,\tau_1^*}$. We repeat this to obtain synthetic values on all records for $Z_1, ..., Z_f$. The illustration in Figure 2 shows the synthetic data set up to this point. Values have been generated using quantile regression predictions and appear in the synthetic set (shaded). Variables $Z_{f+1}, ..., Z_s$ are still empty, or missing.

Finally, we generate values for $Z_{f+1}, ..., Z_s$ using hot deck imputation and rank swapping. Within categories defined by values of variables in X, we compare synthetic values

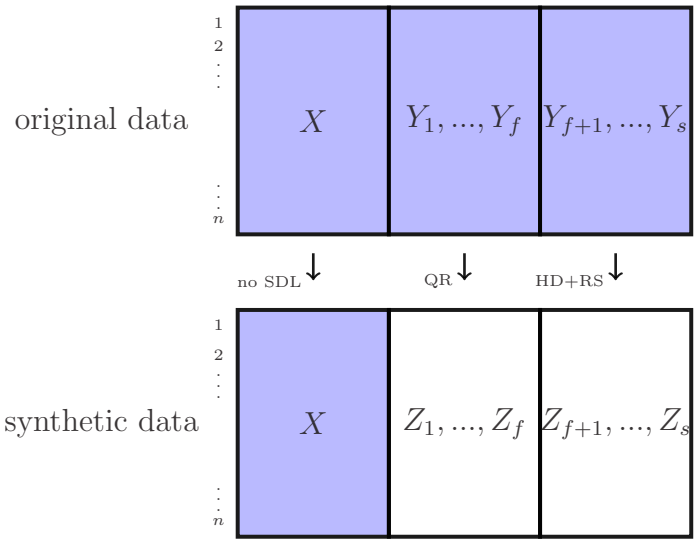

Figure 2.1 An illustration of the original data set and STEP ONE of the creation of the synthetic data set. The original data set contains variables X and Y. In step one, the nonsensitive variables X are copied directly into the synthetic data set.

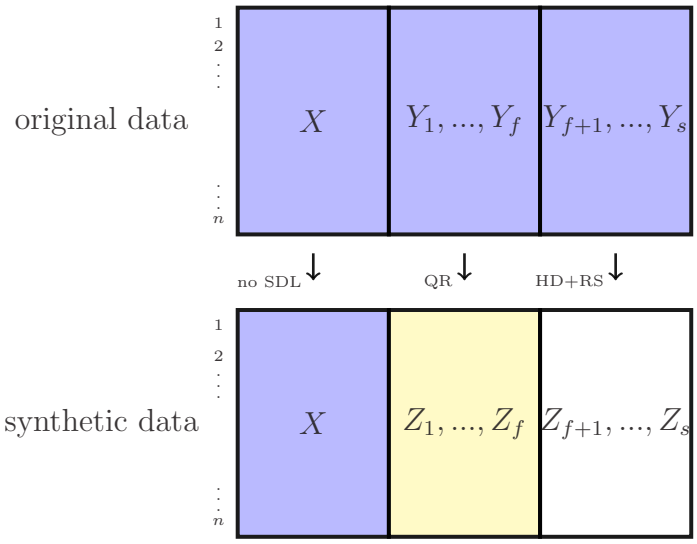

Figure 2.2 An illustration of the original data set and STEP TWO of the creation of the synthetic data set. The original data set contains variables X and Y. In step two, the sensitive variables $Y_1, ..., Y_f$ are replaced in the synthetic data set using quantile regression predictions by variables $Z_1, ..., Z_f$.

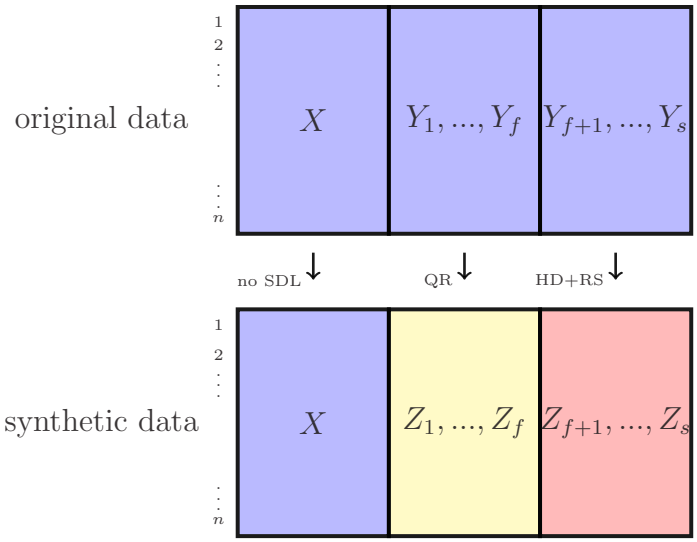

Figure 2.3 An illustration of the original data set and STEP THREE of the creation of the synthetic data set. The original data set contains variables X and Y. In step three, the sensitive variables $Y_{f+1}, ..., Y_s$ are replaced in the synthetic data set using hot deck imputation and rank swapping by variables $Z_{f+1}, ..., Z_s$.

$Z_1, ..., Z_f$ with original values $Y_1, ..., Y_f$ using Mahalanobis distance. Suppose for record j in the synthetic set, we determine record k in the original set has the smallest distance to synthetic record j, i.e. $d(k, j) \leq d(i, j), \quad i = 1, ..., n$. Then for the closest original record k, we compute the sample ranks of each value $y_{f+1}, ..., y_s$ to be $r_{f+1}, ..., r_s$, respectively. To perform rank swapping, we randomly draw ranks $r^*_{f+1}, ..., r^*_s$ from $Uniform(r_{f+1} - \delta_{f+1}, r_{f+1} + \delta_{f+1}), ..., Uniform(r_s - \delta_s, r_s + \delta_s)$, respectively, and impute values with ranks $r^*_{f+1}, ..., r^*_s$. In this way, we fill in the remaining variables in the synthetic set, as illustrated in Figure 3.

In Chapter 3, we study the disclosure risk and data utility of the resulting synthetic data set. In Chapter 4, we present results from three applications.

CHAPTER 3. DISCLOSURE RISK AND DATA UTILITY

Statistical agencies go to great lengths to collect high quality data that are representative of the populations they aim to study. The agencies want to release the data to users for analysis and study. If the data contain identifying variables, such as Social Security number or name and address, then obviously the data set cannot be released without allowing someone to identify the individuals in the data set. In order to release data without violating pledges of confidentiality to respondents, agencies restrict variables that are released and implement statistical disclosure limitation (SDL) methods such as those described in Chapter 2. But which SDL methods should be used? The agencies need to ensure confidentiality protection. They also need to release data that are of value to researchers. Therefore, quantitative measures of disclosure risk and the usefulness of data for inferential purposes are needed. A few authors have proposed measures of the risk of disclosure or violation of confidentiality agreements. These are reviewed in detail in Section 3.1. The extension of these methods to the proposed method of disclosure limitation are explicated in Section 3.2. Three types of intruders, or individuals seeking to identify respondents, are considered. The utility conceptually is the degree to which the data enable a researcher to answer questions. Suggestions for quantifying data utility are presented in Section 3.3.

In general, there is a trade off between reducing disclosure risk and increasing data utility. At one extreme, releasing no data has zero risk (except for someone physically stealing the data), but no usefulness at all. At the other extreme, releasing all collected data, including personal identifying information (or at least everything but explicit personal identifiers), should be the most useful for researchers, but has the highest potential for harm to respondents. Most traditional methods can be compared in terms of the degree to which disclosure protection and data utility are achieved. Researchers at the National Institute of Statistical Sciences and collaborators have called this the R-U Confidentiality Map, where R is a measure of disclosure risk and U is a measure of data utility. Measures R and U are measured for potential data sets, those with the lowest risk and highest utility are candidates to be released. The goal of the research work in this dissertation is to provide a method using

synthetic data generation that greatly reduces the disclosure risk but maintains a reasonably high level of data utility.

3.1 Disclosure risk

In order to judge the relative merits of disclosure limitation methods, we need to assess both the risk of disclosing confidential information and the data utility, or the inferential worthiness, of the released data set. Released data sets that carry too much disclosure risk or too little data utility should be avoided. A review of the literature on measures of disclosure risk and data ultility is presented in this section. Details about the framework developed in Duncan and Lambert (1986, 1989) and Reiter (2005) are presented here. We use and extend existing risk measures to measure disclosure risk for a synthetic data set resulting from our proposed procedure. This is presented in Section 3.2.

3.1.1 Introduction and background

Recall the scenario that was described previously about releasing data on the sex, race, county of residence, and income of survey respondents. In some counties, for some sex and race combinations, there could be very few individuals with extreme income. Based on a release of actual respondent information, someone interested in extorting money or pursuing economic relations with a wealthy person in the survey could perhaps identify the income or other financial details about a respondent that the respondent would not want made public.

In Duncan and Lambert (1986), the authors present assumptions under which they develop a framework for measuring the disclosure risk of a released data set based on assumptions about an intruders behavior. The assumptions are that an intruder can combine information in the released data set with information from external sources to gain information about a *target.* The target can be any respondent to the survey that has data released. They note that almost any data release can provide new information about a target, implying that total avoidance of disclosure is impossible, but that it is possible to control the disclosure to be below an acceptable level. Under these general assumptions, the authors describe a method to quantify the extent of disclosure risk using predictive distributions to model the information an intruder has both before and after data are released. They measure the extent of disclosure risk by comparing characteristics of the predictive distributions

before and after data release.

Duncan and Lambert (1986) propose an indirect probabilistic matching approach to compute the probability of identification, with applications for a target with categorical data or information released in tabular format. Direct probabilistic matching involves linking released records to records in external sources. It necessitates collecting those sources as well as performing the linking procedures and analyzing resulting links. Indirect probabilistic matching, on the other hand, uses the original and released data to compute probabilities of identifying a target in the released data. It does not require external data sources and is flexible to incorporate varying assumptions about the intruder's knowledge, relationships among variables, and disclosure limitation techniques used (Duncan and Lambert 1986, 1989, Reiter 2005). In Duncan and Lambert (1989), the authors focus on matching for continuous data and also expand the decision theoretic framework in applications allowing the flexibility to include an intruder's possible objectives, strategy for compromising the data base, and information gained by the intruder. They distinguish between identity (who the respondent is) and attribute (knowing details about the repondent) disclosure in examples of computing the probability of an intruder linking a record with the target.

Reiter (2005) presents details of working within the Duncan and Lambert framework to measure disclosure risk based specifically on the following scenario. An agency collects data, alters that data using some disclosure limitation method, and releases the altered data for public use. Further, once the data are released, an intruder will attempt to link records in the released data set using information from external sources in order to identify a target respondent. The target represents a respondent (an individual or establishment) whose information is possibly contained on one record in the agency's data. The intruder will consider probabilities of identification to link the target with a record in the released data set based on both information s/he has from external sources and information in the released data set. The record(s) with high enough probability will be considered a match with the target and thus identify the target in the released data set, breaching the promise of confidentiality made to respondents. Reiter (2005) also presents further details and implementation of the Duncan and Lambert approach for data sets altered using traditional disclosure limitation methods such as swapping, recoding and topcoding, noise addition, and combinations of these methods.

We propose to use ideas and extensions thereof presented in Duncan and Lambert (1986, 1989) and Reiter (2005) to measure disclosure risk of a synthetic data set created using pre-

dictions from a conditional model and imputed values. In 3.1.2, we present the framework in Reiter (2005) from which we develop a disclosure risk measure for synthetic data. Section 3.1.3 defines notation. In Section 3.1.4, we present details on probabilistic methods used under various schema of intruder knowledge and disclosure limitation techniques presented in Duncan and Lambert (1986, 1989) and Reiter (2005). We develop extensions to existing risk measures for synthetic data in Section 3.2. These extensions are applicable to our disclosure limitation methodology.

3.1.2 Disclosure risk framework

We develop a measure of disclosure risk under the framework presented in Duncan and Lambert (1986, 1989) and later extended in Reiter (2005). In this framework, we suppose an intruder attempts to identify one or several respondents in the released data. The authors refer to the targeted respondent as a *target.* Unless the intruder has complete and accurate information and the released data are not perturbed sufficiently, the intruder cannot identify the target directly. Instead, the intruder must determine the likelihood of each released record to belong to the target and select the most likely record. Under the Duncan and Lambert framework, this likelihood is measured by combining information the intruder has from an external source prior to any data release with information gained after the data are released using prior and posterior probability distributions.

Predictive probability distributions

Suppose the original data set contains variables Y_1 and Y_2. To compute disclosure risk under the framework being presented here, we aim to compute $Pr(Z_2|Y_1)$, the probabilitiy of observing the released values Z_2 (synthetic versions of values in Y_2) given original values released for variable Y_1. We do so using Bayes' rule; this is described here. In this context, a prior predictive distribution describes the assumptions an intruder makes about values in the original data set. Suppose that prior to any data being released, the intruder knows or assumes that the association between Y_2 and Y_1 can be described well using the linear regression model $Y_2 = \beta_0 + \beta_1 Y_1 + \epsilon$, where ϵ are independently identically distributed random variables from a Normal distribution with zero mean and variance σ_ϵ^2, and that accurate estimates of the regression parameters are available. Then the prior predictive distribution of Y_2 given Y_1 is Normal $(\beta_0 + \beta_1 Y_1, \sigma_\epsilon^2)$.

Now suppose that original values of Y_1 can be released freely to the public but only masked or perturbed values of Y_2 can be released. An SDL method will be used to perturb the values of Y_2 and obtain releasable values for variable Z_2, say. If the intruder has information about the SDL method used to obtain values for Z_2, s/he can assume a posterior predictive distribution of Z_2 given Y_2 and Y_1 which can be used to compute the probability of observing released values on Z_2, given Y_2 and Y_1, or $Pr(Z_2|Y_2, Y_1)$. Recall, our goal is to compute $Pr(Z_2|Y_1)$. We can do so by using Bayes' Rule to write $Pr(Z_2|Y_2, Y_1)$ as

$$Pr(Z_2|Y_2, Y_1) = \frac{Pr(Z_2, Y_2|Y_1)}{Pr(Y_2|Y_1)},$$

where $Pr(Z_2|Y_2, Y_1)$ is the posterior predictive distribution of Z_2, based on the SDL method used and $Pr(Y_2|Y_1)$ is the prior predictive distribution of Y_2. Solving the above for $Pr(Z_2, Y_2|Y_1)$ and integrating out Y_2, we obtain an expression for $Pr(Z_2|Y_1)$:

$$\begin{aligned} Pr(Z_2|Y_1) &= \int Pr(Z_2, Y_2|Y_1)dy_2 \\ &= \int Pr(Z_2|Y_2, Y_1)Pr(Y_2|Y_1)dy_2. \end{aligned}$$

The following example illustrates the use of predictive distributions described above. The model is that Y_2 is related to Y_1 by a normal linear regression model:

$$Y_2|Y_1 \sim N(\beta_0 + \beta_1 Y_1, \sigma_\epsilon^2).$$

Assume that the value of $_2$ is perturbed by adding random noise: $Z_2 = Y_2 + e$, where $e \sim N(0, \gamma^2)$, say. Then $Z_2|Y_2 \sim N(Y_2, \gamma^2)$ and $Z_2|Y_1 \sim N(N(\beta_0 + \beta_1 Y_1, \sigma_\epsilon^2 + \gamma^2)$.

Duncan and Lambert framework

The theoretical framework in Duncan and Lambert (1986) for measuring the extent of disclosure for released data is based on two principles:

1. *The complete state of a user's uncertainty about a target before and after data release is specified by the user's prior and posterior predictive distributions, respectively.*

2. *The user's uncertainty about a target can be summarized by applying a nonnegative concave function $U(\cdot)$ to the user's predictive distribution for the target (DeGroot 1962, 1970). Functions $U(\cdot)$ are called uncertainty functions. The larger the value of U, the more uncertainty.*

The nonnegative concave function $U(\cdot)$ typically will be a function of the predictive probability distribution and a loss function. The loss function is a loss function for the intruder regarding the decisions that the intruder can make.

We assume that the intruder attempts to identify a respondent's record in the released data set. The authors propose the use of uncertainty functions to express characteristics of predictive distributions before and after data are released. By requiring constraints on posterior uncertainty functions, an agency can consider the relative merit of disclosure limitation procedures. Uncertainty reflects the extent of disclosure and is measured with respect to prior and posterior information or knowledge an intruder may possess. As Duncan and Lambert discuss, we can consider posterior knowledge as the knowledge an intruder has based on the released data set. Once the data have been released, the intruder can gain knowledge from the released data, compared to what was known before. This gain is equated to the difference between posterior information and prior information. The knowledge gained from the released data set can also be considered relative to the prior knowledge held before data were released. This is referred to as the relative knowledge gain. The amount of knowledge the intruder has is inversely proportional to the intruder's uncertainty. High uncertainty reflects little knowledge and low uncertainty reflects more knowledge. Higher uncertainty indicates a lower extent of disclosure.

Using these ideas about knowledge and uncertainty, disclosure measures based on uncertainty functions are presented as belonging to one of the following classes: knowledge $U(posterior)$, knowledge gain $U(prior) - U(posterior)$, or relative knowledge gain, $\frac{U(prior)-U(posterior)}{U(prior)}$. For particular prior and posterior predictive distributions, an agency can set a threshold level not to be exceeded by relative information gain. Higher uncertainty after data are released implies higher disclosure limitation, or lower disclosure risk. Methods proposed to be used to protect data sets can be assessed and compared relative to their their disclosure measures.

Duncan and Lambert (1986) suggest chosing an uncertainty function $U(\cdot)$ by considering the information's potential to compromise confidentiality. This is done under a decision-theoretic framework in which the intruder's decision problem is to "identify the target." The target wishes to identify the released record i in Z that contains the same identifying information as the target, t_0, such that $z_{0i} = t_0$. Recall, however, that information on identifying variables t_0 and Y_0 or Z_0 is never released, so the intruder must use available information to make this identification, or link. An intruder can decide to link record i in

the released data set with the target, or decide to not link any record i with the target (null link). If the intruder decides to link the wrong record with the target or if s/he decides to not link any record with the target when the target's record in in fact released, then the wrong decision is made, and the intruder is said to incur a loss. Loss is a function of the decision and the truth, $\mathcal{L}(\cdot)$, where the wrong decision corresponds with a positive value of loss. The decision can be thought of as the value of the link, taking values $z_{i0},\ i = 1, ..., n$ when the intruder decides to link record i with the target and ϕ for the null link.

Duncan and Lambert (1989) use the loss function

$$\begin{aligned} &= 0 \quad \text{if link } = z_{0i} \text{ and } z_{0i} = t_0 \\ \mathcal{L}(link, truth) &= l_1 \quad \text{if link } = \phi \text{ and } t_0 \in Z_0 \\ &= l_2 \quad \text{if link } = z_{0i} \text{ and } z_{0i} \neq t_0. \end{aligned}$$

Zero loss is incurred when a correct link is made. A loss of l_1 is incurred when no link is made, but the target record is released in the data set Z. A loss of l_2 is incurred when an incorrect link is made. Since the goal of the intruder is to identify the target, or make a correct decision or link, s/he also wishes to minimize the loss. If the $truth$ were known to the intruder, the best decision, or optimal $link$ could be chosen to minimize loss. Since the intruder does not know the $truth$, however, s/he can consider minimizing the expected loss over all possible decisions or link values.

Suppose an intruder attempts to identify target t in released data set Z. Suppose Z contains only variable Z_1, then an identification, or link can occur for any value z_1 that the intruder believes to be the released value of the target's variable t_1. If the intruder has some information on the SDL procedure used to produce Z from Y, and some prior information about the distribution of values on Y, then a posterior predictive distribution $p(\cdot)$ can be formulated for the values on Z_1, call it $p_y(z_1) = p(z_1|y)$. Presumably, the target's value on t_1 has this same distribution. Then the expected loss is the expected value of $\mathcal{L}(z_1, d)$ under $p(z_1|y)$, or $\sum \mathcal{L}(z_1, d)p_y(z_1)$ when Z_1 is discrete. To minimize the expected loss, then, evaluate $\inf_d \sum \mathcal{L}(z_1, d)p_y(z_1)$, where d are the possible decisions or links ($link \in \{z_{0i}, \phi\}$ in the example above).

The optimal decision is the value of d that minimizes the expected loss. From the intruder's point of view, the smaller , $\inf_d \sum \mathcal{L}(z_1, d)p_y(z_1)$ is, the more certain s/he can be that the record containing value z_1 belongs to the target. From the agency's perspective, a less certain intruder leads to fewer links, so the agency would actually like $\inf_d \sum \mathcal{L}(z_1, d)p_y(z_1)$,

or uncertainty, to be large. Duncan and Lambert (1986, 1989) equate this expression with the uncertainty function, $U(p) = \inf_d \sum \mathcal{L}(z_1, d) p_y(z_1)$. The agency's goal to dissuade an intruder that links can be acheived by maximizing uncertainty, which can be done by controlling the predictive distribution $p_y(z_1)$ through the SDL procedure. Fewer links can be associated with higher uncertainty which is associated with lower disclosure risk.

In Reiter (2005), the author provides details on measuring disclosure risk under the Duncand and Lambert framework for released data generated using recoding, swapping, and random noise addition. Notation is presented in Section 3.1.3 and details of disclosure risk measurement are presented in Section 3.1.4.

3.1.3 Notation

We use the notation presented in Reiter (2005) for data generated using traditional methods of SDL and extend it to formulate disclosure risk for data generated using the proposed synthetic data method. Suppose an agency collects data on n respondents. Each record in the original data set is associated with one respondent, containing values on several variables. The original data set is referred to as Y, with variables $Y_0, Y_1, ..., Y_d$. The agency uses one to several disclosure limitation methods to perturb data on one or more variables. The perturbed data are released in Z. Data set Z contains $r \leq n$ records with values on variables $Z_1, ..., Z_d$. The variables Y_0 correspond to unique identifiers and are never released in any form.

We assume an intruder with access to the released data will attempt to link records in Z using information available on a target t from external sources. The intruder is assumed to compute the probability that record j in the released data set belongs to the target, conditional on the information t that the intruder has on the target and information contained in Z. This probability is denoted $Pr(J = j|t, Z)$. The larger the probability, the more likely the intruder would declare record j to identify the target. This notation is summarized in Table 3.1 for easy reference in further reading.

The original data set contains records $j = 1, 2, ..., n$, with data on several variables, $k = 0, 1, ..., d$. The agency may release n records or a sample of r records, $r \leq n$. The agency may release d variables or a subset of the variables. Directly identifying information, recorded on variables in $k = 0$, is never released. These variables could include name, social

Table 3.1 Notation for use in disclosure risk formulation.

Notation	Description
Y	original data set
Z	released data set
t	target's original data
$Pr(J = j\|t, Z)$	conditional probability that record j belongs to target t

security number, exact address, etc. Remaining variables, $k = 1, ..., d$, are separated into available and unavailable sets, denoted A and U, respectively. The sets describe the nature of the data with respect to a potential intruder. Specifically, variables in A contain information available from outside sources. Variables in U contain information that is unavailable to an intruder except from the released data. Available variables are further divided into variables that are perturbed Ap (available perturbed) and variables that do not get perturbed Ad (available not perturbed) before being released by the agency. This division could also be done for the variables in U. Further, variables in the released data set that the intruder cannot match with 100% certainty belong to the set C, where $C = (Ap, U)$. All variables in C have been perturbed and/or are not known to the intruder before data are released from the agency.

To clarify and for easy reference in further reading, expanded notation is presented in Table 3.2.

Data on each record j in the original data set Y is denoted y_j, with entries y_{jk} on variables $k = 0, 1, ..., d$. The notation y_j^A is used to denote original data on available variables and y_j^{Ap} to denote data on available perturbed variables in the j^{th} record. Similar notation is used for Ad, U, and C, for data on records in the released data set Z, and in the target's data t. Some variables have equal values by definition. For example, $t^A = y^A$ for all records, since information on the target is assumed to include original data on available variables. Also, $t^{Ad} = y^{Ad} = z^{Ad}$ for all records since variables in the set Ad do not get perturbed. However, even though $t^{Ap} = y^{Ap}$, $t^{Ap} = y^{Ap}$ does not necessarily equal z^{Ap} since variables in the set Ap are perturbed from their original values. In the next section we use this notation to clearly describe the probabilistic framework presented and implemented by Duncan and Lambert (1986, 1989) and Reiter (2005).

Table 3.2 Extended notation for use in disclosure risk formulation.

Notation	Description
Y	original data set
y_{jk}	original data on record j, variable k
Z	released data set
z_{jk}	released data on record j, variable k
t	target's original data
j	record, $j = 1, ..., n$
n	number of original records
r	number of released records, $r \leq n$
k	variable, $k = 0, 1, ...d$
$k = 0$	directly identifying information, never released
A	available variables
U	unavailable variables
Ap	available perturbed variables
Ad	available not perturbed variables
C	variables intruder cannot know with 100% certainty, $C = (Ap, U)$

3.1.4 Probabilistic risk measures

Using the notation and framework described in Sections 3.1.2 and 3.1.3, we procede to discuss the methods proposed in Reiter (2005), based on the framework presented in Duncan and Lambert (1986, 1989). Recall, the agency has collected data set Y, performed some masking or perturbation to protect respondents' confidentiality, and released data set Z. The intruder is assumed to have some information on a target (this can be one or several respondents) available from external sources; this information is contained in t. The intruder seeks to identify the record in Z that belongs to the target. The authors seek to quantify the amount of information an intruder has both before data are released and after data are released in order to compute the probability that the intruder will correctly identify a target in the released data set.

Suppose an intruder has information on target t, and attempts to match record j in Z when the unique identifying information (never released) in t and record j is equal, or when $t_0 = z_{j0}$. The intruder simultaneously attempts to not match target t with record j when

the unique identifying information is not equal, or when $t_0 \neq z_{j0}$. It is assumed that the intruder will compute the probability that record j belongs to the target given the available information on the target, t, and information in the released data set, Z. It is also assumed that t and Z contain some of the same variables, otherwise matching would be a moot point. Using information in t and the released data set Z, the intruder might compute the conditional probability $Pr(J = j|t, Z)$, where J is a random variable taking value j when $t_0 = z_{j0}$ and $r + 1$ when the target record is in the collected data set Y but not in the released data set Z. This could happen when the original data Y is sampled with sample size r strictly less than the number of records in the original data set, n. When $r < n$, $r+1$ represents any of the $n - r$ records in Y not released in Z.

In order to compute $Pr(J = j|t, Z)$, Reiter breaks the probability into manageable parts reflecting information an intruder might have. Such information includes the number of records in the released data set with values on unperturbed variables the same as those values in the target's data, the disclosure limitation methods used to perturb data, prior beliefs about the data on unavailable variables, and multivariate relationships among variables in the data set. Using Bayes' rule, Reiter (2005) expresses the probability of identification as

$$\begin{aligned} Pr(J = j|t, Z) &= Pr(J = j|t, Z^{Ad}, Z^C) \\ &= \frac{Pr(J=j, Z^C|t, Z^{Ad})}{Pr(Z^C|t, Z^{Ad})} \\ &= \frac{Pr(Z^C|J=j,t,Z^{Ad})Pr(J=j|t,Z^{Ad})}{\sum_{j=1}^{r+1} Pr(Z^C|J=j,t,Z^{Ad})Pr(J=j|t,Z^{Ad})}. \end{aligned} \tag{3.1}$$

This follows directly from Bayes' rule if we consider the partition of variables into unavailable, available perturbed, and available unperturbed. Recall, $Z = (Z^U, Z^A) = (Z^U, Z^{Ap}, Z^{Ad})$. So that if we write $Pr(J = j|t, Z) = Pr(J = j|t, Z^U, Z^{Ap}, Z^{Ad})$, and also note that $Z^C = (Z^U, Z^{Ap})$, we can write

$$Pr(J = j|t, Z) = Pr(J = j|t, Z^C, Z^{Ad}). \tag{3.2}$$

Further, using Bayes' rule, we write the conditional probability that $J = j$ given Z^C, t, and Z^{Ad} as the joint probability of $J = j$ and Z^C given t, Z^{Ad} divided by the marginal probability of Z^C given t and Z^{Ad} :

$$Pr(J = j|t, Z^C, Z^{Ad}) = \frac{Pr(J = j, Z^C|t, Z^{Ad})}{Pr(Z^C|t, Z^{Ad})}. \tag{3.3}$$

The numerator in Equation 3 can be decomposed into

$$Pr(J = j, Z^C|t, Z^{Ad}) = Pr(Z^C|J = j, t, Z^{Ad})Pr(J = j|t, Z^{Ad}). \tag{3.4}$$

The marginal distribution of Z^C in the denominator of Equation 3.3 can be computed by summing the joint distribution of Z^C and J over values of J. Since J takes discrete values $1, ..., r+1$ the summation is appropriate here. The expression in Equation 3.4 can also be plugged in for the summand in the denominator of Equation 3.3. Finally, combining terms, we obtain the expression in Equation 3.1.

Computing $Pr(J = j|t, Z)$ using Equation 3.1 becomes a matter of computing its parts for each record j.

1. Component $Pr(J = j|t, Z^{Ad})$

First, consider estimating $Pr(J = j|t, Z^{Ad})$. The variables in Z^{Ad} are released unperturbed, i.e. $Z^{Ad} = Y^{Ad}$. Thus, any record in Z with $z^{Ad} = t^{Ad}$ could potentially be identified as the target's. If $n_t = \#$ records in Z with $z_j^{Ad} = t^{Ad}$, then based only on this information, assuming the intruder knows the target is released in Z, the intruder has a $1/n_t$ chance of correctly identifying target t in Z. In other words, the probability of identifying record j as the target given the target's information and values in Z^{Ad} is

$$Pr(J = j|t, Z^{Ad}) = \frac{1}{n_t}. \tag{3.5}$$

Note, the result in Equation 3.5 depends on the assumption that the intruder knows the target is released in Z, i.e. $Pr(J = r+1|t, Z^{Ad}) = 0$. We assume throughout that the intruder knows the target record is released in Z, that is, $j \leq r$. This is conservative as well as computationally convenient and we assume this throughout the remainder of our discussion on disclosure risk measurement. Details for computing $Pr(J = r+1|t, Z^{Ad}) \geq 0$ are discussed in Reiter (2005).

2. Component $Pr(Z^C|J = j, t, Z^{Ad})$

The probability $Pr(Z^C|J = j, t, Z^{Ad})$ is the probability of observing values on variables Z^C in the released data set given the j^{th} record belongs to the target, the intruder's information on the target, and the values on the variables that are released unperturbed. Variables Z^C are variables that the intruder cannot know with certainty. If we assume the records in the released data set are independent, we can write the joint probability of Z^C as a product of its marginal probabilities, or

$$\begin{aligned} Pr(Z^C|J=j,t,Z^{Ad}) &= Pr(z_1^C,...,z_r^C|J=j,t,Z^{Ad}) \\ &= Pr(z_1^C,...,z_{j-1}^C,z_{j+1}^C,...,z_r^C|J=j,t,Z^{Ad})Pr(z_j^C|J=j,t,Z^{Ad}). \end{aligned} \tag{3.6}$$

Further, if we break Z^C into (Z^U, Z^{Ap}) for the j^{th} record and write the joint probability as a marginal times a conditional, then we have

$$\begin{aligned} Pr(z_j^C|J=j,t,Z^{Ad}) &= Pr(z_j^U,z_j^{Ap}|J=j,t,Z^{Ad}) \\ &= Pr(z_j^U|z_j^{Ap},J=j,t,Z^{Ad})Pr(z_j^{Ap}|J=j,t,Z^{Ad}) \end{aligned} \tag{3.7}$$

This allows us to write

$$\begin{aligned} Pr(Z^C|J=j,t,Z^{Ad}) &= Pr(z_1^C,...,z_{j-1}^C,z_{j+1}^C,...,z_r^C|z_j^C,J=j,t,Z^{Ad}) \\ &\quad \times Pr(z_j^U|z_j^{Ap},J=j,t,Z^{Ad}) \\ &\quad \times Pr(z_j^{Ap}|J=j,t,Z^{Ad}), \end{aligned} \tag{3.8}$$

as expressed in Reiter (2005).

We can incorporate various assumptions about an intruder's knowledge into our computation of the probability of identification and hence into our computation of disclosure risk (Reiter 2005). Each term in the right hand side of $Pr(Z^C|J=j,t,Z^{Ad})$ in Equation 3.8 can be formulated using different assumptions of intruder knowledge and behavior, the particular record an intruder can target, and various assumptions about the variables in the original and released data sets.

Recall that Z^C includes Z^{Ap}–available perturbed, Z^{Up}–unavailable perturbed, and Z^{Ud}–unavailable unperturbed variables. We incorporate knowledge about the SDL method used into $Pr(z_j^{Ap}|J=j,t,Z^{Ad})$ and $Pr(z_j^U|z_j^{Ap},J=j,t,Z^{Ad})$, since these are probabilities of observing released values that have had some SDL method applied to them. We can also incorporate various assumptions about the univariate and multivariate distributions of variables. Formulations presented in Reiter (2005) are presented here. In Section 3.2, formulations based on the proposed SDL method to generate synthetic data using quantile regression and hot deck with rank swapping are introduced, with details on computing probabilities associated with each component.

2a. Component $Pr(z_j^{Ap}|J=j,t,Z^{Ad})$

Recall that Z^{Ap} contains available perturbed variables. Reiter (2005) suggests that if SDL methods are applied to variables independently, then the joint conditional distribution of

z_j^{Ap} given $J = j, t, Z^{Ad}$ is the product of the marginal conditional distributions and the probability can be written as

$$Pr(z_j^{Ap}|J = j, t, Z^{Ad}) = \prod_k Pr(z_{jk}^{Ap}|J = j, t, Z^{Ad}). \tag{3.9}$$

Many SDL methods do generate perturbed values independently from variable to variable. This approach could be taken, for example, using noise addition, with randomly selected errors added to observed values on different variables coming from different distributions. If random swapping of some variable values was performed, this could also be done independently on the variables. Disclosure limitation methods used in the work of Duncan, Lambert, and Reiter include data swapping and adding Gaussian noise. The authors provide examples of formulating expressions for Equation 3.9.

In our proposed SDL method, however, synthetic values based on quantile regression are generated using the conditional sequential models described in Section 2.1.3. Supopse quantile regression is used to generate values on variables Z_1, Z_2, and Z_3 from estimates for the models $Y_1|X$, $Y_2|Y_1, X$, and $Y_3|Y_2, Y_1, X$, respectively. Using estimated quantile regression parameters, values of X are used to generate values for Z_1. Values of X and Z_1 are used to generate values for Z_2. Values of X, Z_1, and Z_2 are used to generate values for Z_3. This obviously implies that the distributions of Z_1, Z_2, and Z_3 are not independent. In the example here, we consider writing the joint probability as the product of marginal probabilities and conditional probabilities:

$$\begin{aligned} Pr(z_j^{Ap}|J = j, t, Z^{Ad}) \;=\; & Pr(z_{1,j}^{Ap}|J = j, t, Z^{Ad}) \\ & \times Pr(z_{2,j}^{Ap}|z_{1,j}^{Ap}, J = j, t, Z^{Ad}) \\ & \times Pr(z_{3,j}^{Ap}|z_{2,j}^{Ap}, z_{1,j}^{Ap}, J = j, t, Z^{Ad}) \end{aligned} \tag{3.10}$$

Details on doing this in the context of our proposed SDL method are presented in Section 3.2.

2b. Component $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$

Consider computing $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$ in Equation 3.8. Recall, variables in U are unavailable, i.e. unknown, to the intruder before data set Z is released. The intruder does not have corresponding values in t for any of the variables in U. The intruder may have prior beliefs about y_j^U values and hence z_j^U. The extent and accuracy of these prior beliefs can vary and resulting probability or risk can be compared among possibilities. If we rewrite

this component of probability as suggested in Reiter (2005), we obtain

$$Pr(z_j^U | z_j^{Ap}, J = j, t, Z^{Ad}) = \int Pr(z_j^U | y_j^U, z_j^{Ap}, J = j, t, Z^{Ad}) Pr(y_j^U | z_j^{Ap}, J = j, t, Z^{Ad}) dy_j^U. \quad (3.11)$$

The second term of the integral reflects prior information on the distribution of variables in U. The intruder can specify prior information based on knowledge from outside sources. Examples in Reiter (2005) include using parameter estimates from a regression of Y_k^U on Z^A, using partial information such as bounds on the target's value of y_{jk}, or even using a uniform distribution over a reasonable range in the case that the intruder has no prior beliefs.

For any variable k in U that is not altered before the data are released, and assuming the intruder knows which variables are not altered, $Pr(z_{jk}|z_j^{Ap}, J = j, t, Z^{Ad}) = 1$. For any variable k in U that is altered, $Pr(z_{jk}|y_{jk}, z_j^{Ap}, J = j, t, Z^{Ad})$ is specified to reflect the disclosure limitation method applied to y_{jk}. This can be done in the same manner described in above from Reiter (2005) or using extensions presented in Section 3.1.3. Reiter offers three suggestions for evaluating the integral in Equation 3.12.

First, we can set the entire probability to 1, which is what we do for a naive intruder. Second, given specifications for the true distribution functions on the reight hand side of Equation 3.12, one can approximate the integral numerically or, if possible, directly evaluate it. For example, if y_j^U is related to z_j^{Ap} through an assumed linear regression equation ($y_j^U = \alpha + \beta Z_j^{Ap} + \epsilon$, $\epsilon \sim N(0, \sigma_y^2)$) and if Z_j^U is related to Y_j^U by additive Gaussian noise ($Z_j^U = y_j^U + e$, $e \sim N(0, \sigma_z^2)$), then, the distribution of Z_j^U is $N(\alpha + \beta Z_j^{Ap}, \sigma_y^2 + \sigma_z^2)$. Of course, one needs to know or have an estimate for σ_y^2 and σ_z^2. Third, one can approximate the integral by drawing valaues of y_j^U from $Pr(y_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$, plugging those values into the integral for a given value of Z_j^U, and taking an average. This method requires distributional assumtptions as well.

2c. Component $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad})$

Variables $z_i^C, i \neq j$ include variables that are available to the intruder but are perturbed and variables unavailable to the intruder on units that are not the target. Assuming independence among pairs (z_i^C, y_i^C) across records, $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad})$ is rewritten

$$\left(\prod_{i=1}^{j-1} \int Pr(z_i^C | y_i^C, z_i^{Ad}) Pr(y_i^C | z_i^{Ad}) dy_i^C \right) \left(\prod_{i=j+1}^{r} \int Pr(z_i^C | y_i^C, z_i^{Ad}) Pr(y_i^C | z_i^{Ad}) dy_i^C \right),$$

as in Reiter (2005).

Note that assuming independence across records eliminates conditioning terms z_j^C in the above probabilities corresponding to records $i = 1, ..., j-1, j+1, ..., r$. Further, when Equation 3.13 is plugged into Equation 3.1, the entire term simplifies to

$$1/\int Pr(z_j^C|y_j^C, z_j^{Ad})Pr(y_j^C|z_j^{Ad})dy_j^C)$$

(Reiter 2005). The author again suggests procedures to approximate the integral in the denominator of this expression. As before, one option is to replace the integral with a value of 1. This is expedient, but likely losses some mathcing information. A second option is numerical approximation or, if possible, direct approximation. As before, one must choose distributions and values of some parameters. A third option involves drawing values of Y_j^C from $Pr(y_j^C|z_j^{Ad})dy_j^C)$ and plugging those values in into the integral for a given value of Z_j^C, and taking an average. This method also requires distributional assumptions. In most data sets, the assumption of independence across records is reasonable. Most samples are chosen randomly from a population or some subpopulation to obtain accurate information about the population as a whole.

Summary

The framework and methods presented in Duncan and Lambert(1986, 1989) and Reiter (2005) provide a framework in which to compute the probability of identifiying an individual record. As discussed in Section 3.1.2, an agency wants to dissuade linking any records with any target. This can be acheived if the intruder's uncertainty about the target in teh released data set is high, i.e., if the probability of identifying any record as the target's is low. The agency controls this through the SDL method which influences the posterior predictive distribution of the target's values. Under this framework, disclosure risk is inversely proportional to uncertainty, where high uncertainty is associated with lower risk. In fact, disclosure risk is equated with the probability of identification, where a low probability of identifying any record as the target's implies a low disclosure risk. The probability of identification includes the following two components:

1. $Pr(J = j|t, Z^{Ad})$
2. $Pr(Z^C|J = j, t, Z^{Ad})$

The second component is further divided into three components:

2a. $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$

2b. $Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$

2c. $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad})$

Implications of the SDL method used and any distributional assumptions are incorporated into the measure of disclosure risk through these components.

Whole data set risk

Under this framework, we can consider various levels of intruder knowledge and possible behavior. We can compare disclosure risk between disclosure limitation methods and between records or targets with different levels of sensitivity. Using the probability of identification for every record, we can measure the disclosure risk for the entire released data set. To obtain a disclosure risk measure for the entire data set, Reiter (2005) assumes the intruder has correct records for all units. This is a very conservative assumption, but that might be necessary to provide adequate protection. The intruder is assumed to compute the probability of identification for each record against all records in the original data set. If the agency has determined a level or threshold of risk that any one record, or target, must have risk less than this level, then the disclosure risk for the entire data set can be equated with the number of records with disclosure risk greater than this threshold, the expected number of true matches, and the number of units with unique true matches. We will not focus on computing this for a synthetic data set generated using our proposed method. Such a measure of disclosure risk for the entire data set is valuable for comparisons of methods.

In Section 3.2, we extend their work to assess disclosure risk associated with synthetic data sets under this framework. We discuss details of assessing disclosure risk associated with our proposed procedure of generating synthetic values using quantile regression predictions and hot deck combined with rank swapping. We also suggest similar extensions that can be used to measure disclosure risk associated with synthetic data generated using models with known or estimated distributional properties.

3.2 Disclosure risk for the proposed synthetic data method

In Section 3.1, we presented the framework for measuring disclosure risk as developed in Reiter (2005). Disclosure risk in this framework is equated with the probability of identification. That is, an intruder might compute the probabilities that released records match the target in order to identify a respondent's record in the released data. Previous formulations for the components of this probability, or risk, are based on traditional statistical disclosure limitation techniques such as re-coding, swapping, and noise addition among others. In this section, we develop extensions to measure disclosure risk associated with synthetic data generated using the proposed statistical disclosure limitation procedure presented in Chapter 2 of this dissertation. The proposed method combines quantile regression predictions on randomly selected quantiles with hot deck imputation and rank swapping to produce record level synthetic data for release. In Section 3.2.1, we present the intruder's decision tree and introduce notation. In Section 3.2.2, we describe the formulation of each component of the identification probability. In Section 3.2.3, we describe possible formulations of each component to evaluate disclosure risk for synthetic data generated using our proposed method.

3.2.1 Intruder knowledge and decisions

In this section, we develop formulations for the components of the probability of identification, $Pr(J = j|t, Z)$, described in Section 3.1. We evaluate this probability under different assumptions about intruder behavior and knowledge. We consider intruder behavior to reflect certain decisions an intruder might make based on knowledge of the original data, released data, and disclosure limitation method used to generate the released data. A quite sophisticated intruder will have knowledge about the disclosure limitation method, whereas a less sophosticated intruder will have some knowledge of the data prior to its release. A much less sophisticated intruder will only have knowledge about the data after it is released. Based on the level of knowledge an intruder posesses, s/he can make one of several decisions. Each decision corresponds to an approach the intruder can take to evaluate the probability of identification in the released data corresponding to a particular target. We state the possible decisions with respect to intruder knowledge in a diagram similar to a decision tree. Assuming an intruder wishes to compute the probability of identification for each record, we can follow one of several paths to determine how s/he might do so.

We examine intruders' decisions under varying degrees of knowledge. An intruder's knowledge and behavior lies somewhere between naive and quite sophisticated. We as-

sume the *naive* intruder is one whose knowledge is based on *posterior* information only. This intruder knows nothing about the data except from what s/he learns from the released data. We assume the sophisticated intruder has accurate information about the data before release (*prior* information), information after release (*posterior*) information, and details about the statistical disclosure limitation procedure method used (*SDL* information). We call this the *SDL* intruder. We call the *average* intruder one who combines information s/he has before release–*prior* information, and after release–*posterior* information. Some decisions will be unavailable to an intruder based on assumptions about their knowledge. For example, a naive intruder cannot incorporate knowledge about the SDL methods to evaluate the probability of identification because s/he does not have this knowledge. On the other hand, an intruder with quite accurate details about the data set and SDL methods might choose simpler approaches to compute the probability out of convenience.

Recall from Section 3.1.2, the probability of identification $Pr(J = j|t, Z)$ can be written as the product of four major components relating to available and unavailables variables that are perturbed or remain unperturbed. These components can be seen in Equation 3.1 and Equation 3.8, rewritten below for convenience. To evaluate each component, we consider possible ways an intruder would compute each based on assumptions about his/her knowledge. The goal is to develop a measure that can be applied to the released data set to measure its disclosure risk. This measure should be flexible to incorporate different levels of intruder knowledge and behavior as well as different disclosure limitation methods used to produce the data for release.

Recall from Section 3.1.2,

$$Pr(J = j|t, Z) = \frac{Pr(Z^C|J = j, t, Z^{Ad})Pr(J = j|t, Z^{Ad})}{\sum_{j=1}^{r+1} Pr(Z^C|J = j, t, Z^{Ad})Pr(J = j|t, Z^{Ad})}$$

and

$$\begin{aligned} Pr(Z^C|J = j, t, Z^{Ad}) &= Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad}) \\ &\quad \times Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad}) \\ &\quad \times Pr(z_j^{Ap}|J = j, t, Z^{Ad}). \end{aligned}$$

We refer to the components on the right-hand-side of $Pr(Z^C|J = j, t, Z^{Ad})$ as A, B, C, respectively, where

$$
\begin{aligned}
A &= Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad}), \\
B &= Pr(z_j^U | z_j^{Ap}, J = j, t, Z^{Ad}), \text{ and} \\
C &= Pr(z_j^{Ap} | J = j, t, Z^{Ad}).
\end{aligned}
$$

As described above, we characterize intruder knowledge and behavior as *naive*, *average*, or *SDL*. The *naive* intruder is one who only possesses *posterior* information, or information available from the released data. S/he decides to use the posterior information to compute the probability of identification. At the other extreme, an *SDL* intruder is one who has accurate and fairly detailed knowledge of the statistical disclosure limitation procedure used and decides to use this information to compute each component. The *SDL* intruder represents a worst case scenario from an agency's perspective. In between these two extremes is the *average* intruder who has some knowledge of the data prior to its release. The *average* intruder might compute the probability of identification based on *prior* and *posterior* knowledge, but with limited or no knowledge of the disclosure limitation methods used. In addition to *naive, average,* and *SDL* intruders, we also consider a *super naive* intruder, who computes identification probability using only information on the target and on released, unperturbed variables. The *super naive* intruder computes $Pr(J = j|t, Z^{Ad})$ but sets $Pr(Z^C|J = j, t, Z^{Ad})$ equal to 1.

Note that using this framework affords us flexibility in various assumptions about an intruder's behavior with respect knowledge. For example, the *naive* intruder with only *posterior* knowledge can make one of two decisions about how to compute the probability of identification. S/he can compute $Pr(J = j|t, Z^{Ad})$ and use formulations of A_{naive}, B_{naive}, C_{naive} reflecting *posterior* knowledge to obtain a value for $Pr(Z^C|J = j, t, Z^{Ad}) \neq 1$. Alternatively, the intruder can decide to compute $Pr(J = j|t, Z^{Ad})$ combined with the *super naive* decision to set $Pr(Z^C|J = j, t, Z^{Ad}) = 1$. The *SDL* intruder can combine the value for $Pr(J = j|t, Z^{Ad})$ with A_{SDL}, B_{SDL}, C_{SDL} reflecting the SDL method used, or combine $Pr(J = j|t, Z^{Ad})$ with any other formulation of A, B, and C, including the super naive approach to set these components to 1. The intruder with more knowledge can make several possible decisions about how to procede, while the intruder with less knowledge has fewer options.

Table 3.3 provides a summary of the framework under which disclosure risk is computed. Details about levels of information that can be used to formulate each component, A, B, and C, are discussed in the following section. Formulation of the components under varying assumptions about the intruder's knowledge and behavior are presented in Sections 3.2.3 through 3.2.6.

Table 3.3 Intruder knowledge and decisions for disclosure risk formulation.

Goal $= Pr(J = j|t, Z)$

$$Pr(J = j|t, Z) = Pr(J = j|t, Z^{Ad})Pr(Z^C|J = j, t, Z^{Ad})$$

2 main components

$Pr(J = j|t, z^{Ad}) = 1/n_t$

and

$Pr(Z^C|J = j, t, Z^{Ad}) = 1$ super naive

$Pr(Z^C|J = j, t, Z^{Ad}) = ABC$ others

3 intruder types

naive: posterior knowledge
$A_{naive}, B_{naive}, C_{naive}$

average: prior and posterior knowledge
$A_{avg}, B_{avg}, C_{avg}$

SDL: prior and posterior knowledge, SDL method known
$A_{SDL}, B_{SDL}, C_{SDL}$

$A = Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C|z_j^C, J = j, t, Z^{Ad})$
$B = Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$
$C = Pr(z_j^{Ap}|J = j, t, Z^{Ad})$

3.2.2 Component formulation

Formulations for each component of disclosure risk are discussed in this section. To clearly convey how this measure of disclosure risk can be applied to a perturbed or synthetic data set, we descirbe a generic data set. Our proposed SDL procedure is applied to this data set to generate synthetic values. Using this hypothetical scenario, we present details of each component.

Consider data set Y containing variables $Y_1, ..., Y_7, Y^{Ad}$, where Y^{Ad} are the nonsensitive demographic variables, Y^{Ap} are the available perturbed variables, and Y^U, are the unavailable variables. Each of $Y_1, ..., Y_7$ belongs to either Y^{Ap} or Y^U. Table 3.4 shows the nature of each hypothetical variable.

Table 3.4 Variable types in hypothetical data set.

	original	synthetic	
$Y^{Ad}:$	Y^{Ad}	$Z^{Ad} = Y^{Ad}$	
$Y^{Ap}:$	Y_1, Y_2, Y_3, Y_4	$Z_1^{Ap}, Z_2^{Ap},$	(QR)
		Z_3^{Ap}, Z_4^{Ap}	(HD+RS)
$Y^{Up}:$	Y_5, Y_6	Z_5^{Up}, Z_6^{Up}	
$Y^{Ud}:$	Y_7	$Z_7^{Ud} = Y_7^{Ud}$	

Using our proposed SDL procedure, we produce Z for release. Specifically,

1. variables in Z^{Ad} remain unperturbed (possibly re-categorized), i.e. $Z^{Ad} = Y^{Ad} \Rightarrow t^{Ad} = Z_t^{Ad}$, where Z_t^{Ad} is the target's record in the released data,
2. values for z_1^{Ap} are generated conditional on Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e.

$$Q_{z_{1j}^{Ap}}(\tau^*|Z^{Ad}) = Z_j^{Ad}\beta_1(\tau_{1j}^*)\epsilon_1,$$

3. values for z_2^{Ap} are generated conditional on y_1^{Ap}, z_1^{Ap}, and Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e.

$$Q_{z_{2j}^{Ap}}(\tau^*|Z^{Ad}, y_1^{Ap}) = \begin{pmatrix} Z_j^{Ad} \\ z_{1j}^{Ap} \end{pmatrix} \beta_2(\tau_{2j}^*) + \epsilon_2,$$

4. values for z_3^{Ap} and z_4^{Ap} are generated using hot deck imputation and rank swapping, matching based on Mahalanobis distance between synthetic and original values of Z_1 and Z_2

5. values for z_5^{Up} are generated conditional on y_1^{Ap}, z_1^{Ap}, y_2^{Ap}, z_2^{Ap}, and Z^{Ad} using quantile regression predictions at randomly selected quantiles, i.e.

$$Q_{z_{5j}^{Ap}}(\tau^*|Z^{Ad}, y_1^{Ap}, y_{2j}^{Ap}) = \begin{pmatrix} Z_j^{Ad} \\ z_{1j}^{Ap} \\ z_2^{Ap} \end{pmatrix} \beta_5(\tau_{5j}^*) + \epsilon_5,$$

6. values for z_6^{Up} are generated using hot deck imputation and rank swapping, and

7. and values for z_7^{Ud} are left unperturbed in the released data.

When conditioning is used in the proposed SDL procedure, this information can be incorporated to compute the probability of identification reflecting the dependence and independence between variables in Z, resulting from the SDL procedure. If the proposed procedure yields Z with high data ultility, these relationships are reflected in the original data Y. Obviously, quantile regression predictions, z_1, z_2, and z_5, are dependent on conditioning variables in their respective models. In the hot deck procedure we identify matching records in the original data set in order to impute values of z_3, z_4, and z_6, after further rank swapping. Matching is based on the distance between z_1, z_2 in each record and all records y_1, y_2. With a match identified in record i say, values on y_{3i}, y_{4i}, y_{6i} are ranked, and ranks are swapped with randomly selected ranks within some interval with center at the original rank and width $2\delta_k$. Values from the record with the swapped rank are imputed into Z. Swapping occurs independently of other variables, except perhaps Z^{Ad} when swapping is performed within categories. This implies that z_3, z_4, and z_6 are dependent on z_1, z_2, Z^{Ad} but are independent of one another.

First, we describe the formulation of $Pr(J = j|t, Z^{Ad})$ as presented in Reiter (2005). Then we present details on formulating the components of disclosure risk or probability of

identications ($A,\ B$, and C) for an SDL, *average*, and *naive* intruder.

Formulation of $Pr(J = j|t, Z^{Ad})$

For each type of intruder, we use the same formulation for $Pr(J = j|t, Z^{Ad})$ as presented in Reiter (2005). Since this component of the probability is conditioned on information in t and Z^{Ad} only, it is straightforward to set

$$Pr(J = j|t, Z^{Ad}) = \frac{1}{n_t}, \tag{3.12}$$

where n_t is the number of records in Z with $z_j^{Ad} = t^{Ad}$.

3.2.3 Formulation of components for the SDL intruder

Details about possible formulations for the components that go into computing the probability of identication, or disclosure risk, for the SDL intruder are presented here. The components are $A,\ B$, and C. In this section, assumptions and decisions an SDL intruder can make are examined. The resulting forms of $A_{SDL},\ B_{SDL}$, and C_{SDL} are presented. The SDL intruder is the most sophisticated intruder that we consider.

C: Formulation of $C_{SDL} = Pr(z^{Ap}|J = j, t, Z^{Ad})$

Under the SDL procedure outlined in Section 3.2.2, $z^{Ap} = (z_1^{Ap}, z_2^{Ap}, z_3^{Ap}, z_4^{Ap})$ and denotes variables that are available to the intruder and are perturbed before release. Variables z_1^{Ap} and z_2^{Ap} are quantile regression predictions computed conditional on Z^{Ad}. Variables z_3^{Ap} and z_4^{Ap}) are hot deck imputations with rank swapping applied, matching is based on the Mahalanobis distance between $(z^{Ad}, z_1^{Ap}, z_2^{Ap})$ and $(y^{Ad}, y_1^{Ap}, y_2^{Ap})$.

An SDL intruder is assumed to know these details. We assume the SDL intruder uses this knowledge to compute components of the probability of identification for each record. To formulate $C_{SDL} = Pr(z_j^{Ap}|J = j, t, Z^{Ad})$ using knowledge about the SDL method used, the joint probability of observing z_1, z_2, z_3, z_4, given that the j^{th} record belongs to the target and information in t and Z^{Ad}, is written as follows

$$\begin{aligned} C_{SDL} &= Pr(z_j^{Ap}|J=j,t,Z^{Ad}) \\ &= Pr(z_{1j},z_{2j},z_{3j},z_{4j}|J=j,t,Z^{Ad}) \\ &= Pr(z_{1j}|J=j,t,Z^{Ad}) \\ &\quad \times Pr(z_{2j}|z_{1j},J=j,t,Z^{Ad}) \\ &\quad \times Pr(z_{3j}|z_{2j},z_{1j},J=j,t,Z^{Ad}) \\ &\quad \times Pr(z_{4j}|z_{2j},z_{1j},J=j,t,Z^{Ad}), \end{aligned} \tag{3.13}$$

for every j^{th} released record $j = 1, ..., r$. The variable z_{3j} does not appear in the last line because z_3 was not used in the generation of z_4.

In the following paragraphs, we evaluate Equation 3.13 by considering portions of the SDL procedure relevant to each of the right hand side terms. For terms involving z_1 and z_2, probabilitiy measures are based on quantile regression inference results. For terms involving z_3 and z_4, probability measures are based on hot deck imputation and rank swapping.

C1: z_1^{Ap} and z_2^{Ap}

The variables z_1^{Ap} and z_2^{Ap} are the variables that are available to the intruder and are perturbed before release using the regression quantile method. Initially, consider values for z_1 to be generated using quantile regression predictions from model estimates for a single quantile τ. Using this simplification, we develop an expression for the probability, then expand it to formulate an expression to incorporate unknown randomly selected quantiles used in the proposed procedure. Results in Koenker (2005) indicate that, similar to least squares regression, the regression parameter estimates at quantile τ are asymptotically Normal, centered at the true parameter value, with variance dependent on τ. Assumming independent and identically distributed errors, i.e. $\sqrt{n}(\hat{\beta}_n(\tau)-\beta(\tau)) \tilde{\rightarrow} N(0,\omega^2(\tau))$, where $\omega^2(\tau) = \frac{\tau(1-\tau)}{f^2(F^{-1}(\tau))}$. For practical purposes, we assume $\hat{\beta}_n \sim N\left(\beta(\tau), \frac{\omega^2(\tau)}{n}\right)$, which implies $X\hat{\beta}_n \sim N\left(X\beta(\tau), \frac{\omega^2(\tau)}{n}X'X\right)$. For details, see Section 2.1 of this dissertation and Koenker (2005).

If the agency generates values using predictions

$$z_{1j} = \hat{y}_{1j,\tau_{1j}} = z_j^{Ad}\hat{\beta}_1(\tau_1),$$

which the intruder is aware of, and the intruder knows the value of τ_{1j} for every $j = 1, ..., r$, then s/he can formulate $Pr(z_{1j}|J=j,t,Z^{Ad})$ for record $j = 1, ..., r$ to be:

$$\begin{aligned} Pr(z_{1j}|J=j,t,Z^{Ad}) &= \phi_{1j,\tau_{1j}} \\ &= \phi\left(\frac{z_{1j}-t_{1j}}{\omega(\tau_{1j})\sqrt{\frac{1}{n}z_j^{Ad^T}z_j^{Ad}}}\right). \end{aligned} \tag{3.14}$$

We could also assume the intruder knows less. Suppose that s/he knows that the agency generated z_{1j} at one quantile for every j, but does not know the value of $\tau_{1j} = \tau_1$. The intruder might decide to estimate τ_1 or set it equal to some constant. To estimate τ_1, the intruder could use the released data to fit the quantile regression model $z_{1,\tau_1} = Z^{Ad}\hat{\beta}(\tau_1)$ at several values of $\tau_1 = \tilde{\tau}$, compute predicted values $\hat{z}_{1,\tilde{\tau}_1}$ at each $\tilde{\tau}_1$, and compare the resulting predictions with the values of z_1 in the released data. The intruder's estimate of τ_1, $\hat{\tau}_{1,intruder}$ say, might be the value at which the intruder's predicted values differ the least from released values z_1, according to some measure of closeness. If the intruder chooses to set τ_1 equal to some constant, c, such as the median, then $\hat{\tau}_{1,intruder} = c$. The formulation for $Pr(z_{1j}|J=j,t,Z^{Ad})$ for each record $j = 1,...,r$ is simply the expression in Equation 3.14, with τ_{1j} replaced by $\hat{\tau}_{1,intruder}$.

We extend these ideas to compute $Pr(z_{1j}|J=j,t,Z^{Ad})$ for z_{1j} generated at randomly selected quantiles τ_{1j}^*, $j = 1,...,r$. First suppose the intruder knows the values of τ_{1j}^* for each record $j = 1,...,r$, then suppose the intruder estimates τ_{1j}^* or sets it equal to some constant, c_j.

If the intruder knows the value of randomly selected τ_{1j}^* for each $j = 1,...,r$, then $Pr(z_{1j}|J=j,t,Z^{Ad})$ can be expressed as in Equation 3.14 by replacing τ_{1j} with τ_{1j}^*.

If the intruder does not know the values of τ_{1j}^* for each j, s/he can estimate it or set it equal to c_{1j} for each $j = 1,...,r$. The intruder might estimate τ_{1j}^* by fitting the quantile regression model at several quantiles, computing predicted values for each record at each quantile, and comparing z_{1j} with his/her predicted values. The intruder might set $\hat{\tau}_{1j,intruder}^*$ equal to the quantile at which his/her predicted value is closest to z_{1j} on each record. Note that each record would be associated with one (possibly distinct) value of $\hat{\tau}_{1j,intruder}^*$, estimated or set equal to some constant, c_{1j}. The intruder's estimated quantile $\hat{\tau}_{1j,intruder}^*$ can be substituted into Equation 3.14 for τ_{1j}. In general, we write $Pr(z_{1j}|J=j,t,Z^{Ad})$ as in Equation 3.14, with

$$\tau_{1j} = \begin{cases} \tau_{1j}^* & \text{intruder knows value of randomly drawn } \tau_{1j}^* \text{ used to generate } z_{1j} \\ \hat{\tau}_{1j,intruder}^* & \text{intruder estimates value of randomly drawn } \tau_{1j}^* \text{ used to generate } z_{1j}. \end{cases}$$

Though the estimation of τ_{1j} described above is possible, it may not be highly likely that an intruder will expend the time and energy required to fit numerous quantile regression models and make comparisons between the resulting predictions and the released values of z_1. Depending on the size of the released data set and the number of variables in Z^{Ad}, doing so may be practically impossible, depending on computing resources. It would be interesting to implement this approach to determine if it yields accurate and timely estimates of τ_{1j} that in turn produce high probability estimates for the target's released record and low probability estimates for other records, allowing the intruder to identify the target in the released records.

For the application in Section 4.3, we act as if the value of τ_{1j} is known and equal for all records, $\tau_{1j} = \tau$ for every $j = 1, ..., n$. Based on this, the probability of identification is estimated.

Since values for z_2 are also generated using predictions from a quantile regression model, we use the same ideas to formulate $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$. The quantile regression model is:

$$z_{2j} = \hat{y}_{2j,\tau_2} = \begin{pmatrix} Z^{Ad} \\ z_{1j} \end{pmatrix} \hat{\beta}_2(\tau_2).$$

Thus, the formulation for $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$ is also based on the asymptotic Normal distribution of quantile regression parameters. For the intruder's value of τ_{2j},

$$\begin{aligned} Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad}) &= \phi_{2j,\tau_{2j}} \\ &= \begin{cases} \phi\left(\frac{z_{2j}-t_2}{\omega(\tau_{2j})\sqrt{\frac{1}{n}(Z_j^{Ad}\ z_{1j})(Z_j^{Ad}\ z_{1j})^T}}\right), & t^{Ad} = z_j^{Ad} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.15}$$

for every $j = 1, ..., r$, where

$$\tau_{2j} = \begin{cases} \tau^*_{2j} & \text{intruder knows value of randomly drawn } \tau^*_{2j} \text{ used to generate } z_{2j} \\ \hat{\tau}^*_{2j,intruder} & \text{intruder estimates value of randomly drawn } \tau^*_{2j} \text{ used to generate } z_{2j}. \end{cases}$$

C1: Alternative for z_1^{Ap} and z_2^{Ap}

In the Equations 3.14 and 3.15 for conditional probabilites of z_1^{Ap} and z_2^{Ap}, we were willing to assume the approximate asymptotic normality of quantile regression parameter estimates.

This is generally acceptable in our applications, since the regression estimates are based on a large number of records (over 10,000 and up to millions). Conditions D1, D2, and F in Section 2.1.2 might not hold if the data base contains a small number of records or if too many records contain all zeros or very small values. In this case, the knowledgable intruder would not use the asymptotic normal distribution to compute $Pr(z_{1j}|J = j, t, Z^{Ad})$ and $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$.

Alternatively, the target's t_1 and t_2 predicted values, $\hat{t}_1$ and $\hat{t}_2$, could be computed using quantile regression estimates from the released data set and the target values of t^{Ad}. The intruder could then compare the target's predicted values $\hat{t}_1$ and $\hat{t}_2$ to values z_1 and z_2 released in Z. Among records with equal available and unperturbed variable values, $t^{Ad} = Z^{Ad}$, it would be reasonable to consider identifying the target with any record containing z_1 and z_2 values within some range of $\hat{t}_1$ and $\hat{t}_2$. If the intruder is willing to consider any records within an amount $\gamma_1 > 0$, say, of $\hat{t}_1$, then all records within this distance to $\hat{t}_1$ have equal probability of belonging to the target. If the intruder simultaneously considers only records with values of z_2 within $\gamma_2 > 0$ of $\hat{t}_2$, then only the records with values z_1 and z_2 within the intervals $(\hat{t}_1 \pm \gamma_1)$ and $(\hat{t}_2 \pm \gamma_2)$ are considered as potential matches with the target. Records with z_1 and z_2 values within these intervals have equal probability of belonging to the target. Suppose there are n_{t_1,t_2} such records, then the joint probability of observing z_1^{Ap} and z_2^{Ap} conditional on the j^{th} record belonging to the target, the information in t, and the values in Z^{Ad} can be formulated as

$$\begin{aligned} Pr(z_{1j}, z_{2j}|J = j, t, Z^{Ad}) &= Pr(z_{1j}|J = j, t, Z^{Ad}) Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad}) \\ &= \frac{1}{n_{t_1,t_2}} \text{ if } t^{Ad} = Z^{Ad},\ z_1 = \hat{t}_1 \pm \gamma_1,\ z_2 = \hat{t}_2 \pm \gamma_2. \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3.16}$$

In our applications, Conditions D1, D2, and F all seem reasonable due to the large number of records in potential applications, so we use the formulations in Equations 3.14 and 3.15 to compute the probability of identification in our applications. Details and results of implementing this are presented in Section 4.3.

C2: z_3^{Ap} and z_4^{Ap}

The variables z_3^{Ap} and z_4^{Ap} are the variables that are available to the intruder and are perturbed using hot deck imputations with rank swapping before release. We can formulate $Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ and $Pr(z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ based on details of the hot deck and rank swapping procedures. Recall that in the proposed hot deck procedure, we identify potential matches in the original data for each record in the synthetic data. Matching is based on distance between values on variables for which we computed quantile regression predictions. In our example, hot deck is used to identify the closest record i in Y to each record in Z such that distance $d\left[\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix}, \begin{pmatrix} z_{1j} \\ z_{2j} \end{pmatrix}\right]$ is minimized. The sample ranks of y_{3i} and y_{4i}, r_{3i} and r_{4i}, respectively, are computed and random ranks are drawn from discrete Uniform distributions centered at r_{3i} and r_{4i}, each with width $2\delta_k$, $\delta_k > 0$.

With this knowledge of the SDL procedure, the intruder is aware that values z_{3j} and z_{4j} in record j are not likely imputed from the same original record. S/he is aware, however, that imputation begins with computing the Mahalanobis distance between y_{1i}, y_{2i} and z_{1j}, z_{2j}. Initially, it would seem that through replicating the distance calculations, the intruder could identify the record from which values for z_3 and z_4 were swapped from. Here, we discuss why this is not necessarily the case.

Suppose the intruder replicates the hot deck procedure by computing the Mahalanobis distance between the target values t_1 and t_2 and values z_{1j} and z_{2j} in each record using the distance to obtain $d(t, j)$ for each released record. Each distance $d(t, j)$ is the distance the agency computes between the synthetic candidate record j and the donor that is the target's record. This may not be among the original records with the smallest distances to candidate j. In fact, the target record may not have been identified as the closest donor to any of the original records, i.e., even if $d(t, j)$ is the smallest distance among all distances $d(t, j)$, $j = 1, ..., r$ in the released data, $d(t, j)$ may not have been among the smallest distances $d(i, j)$ for synthetic candidate record j and all donors $i = 1, ..., n$ in the original data set. If $d(t, j)$ was not among the smallest $d(i, j)$ for candidate record j and donors $i = 1, ..., n$, the target record would never have been selected for imputation using the hot deck procedure. Values t_3 and t_4 could have been imputed into the synthetic set through the rank swapping procedure if those values had ranks within the range δ_3 and δ_4 of the ranks of values y_3 and y_4 in the matching record.

This leads us to formulate the probability of observing z_3 and z_4 given $z_{2j}, z_{1j}, J = j, t$, and Z^{Ad} independent of the distance between the target values and the values in Z. The

argument above implies that the probability of t_3 and t_4 being imputed into the released data set can be based on the rank swapping portion of the procedure alone. Recall, the probability $Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad})$ is conditional on the j^{th} record belonging to the target. If we assume that the j^{th} record belongs to the target, then we can assume the values t_3, t_4, were swapped with values y_3, y_4 having ranks that were randomly selected from a Uniform distribution over the intervals $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$, respectively. Therefore, for observations z_{3j} and z_{4j} to have been imputed, the ranks in the original record r_{3i} and r_{4i} must fall in the intervals, $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$, respectively. If we assume values in Z_3 and Z_4 have approximately the same ranks as values in Y_3 and Y_4, then the ranks of values z_{3j} and z_{4j}, r^*_{3j} and r^*_{4j}, are also in that interval. We can formulate the conditional probability of observing the z_{3j} and z_{4j} to be equal to the probability of selecting their ranks r^*_{3j} and r^*_{4j} from the intervals $(r_{t_3} - \delta_3, r_{t_3} + \delta_3)$ and $(r_{t_4} - \delta_4, r_{t_4} + \delta_4)$ and 0 for ranks not in these intervals. This can be written

$$\begin{aligned} Pr(z_{3j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) &= \frac{1}{2\delta_3} \text{ if } t^{Ad} = z_j^{Ad} \text{ and } \; r^*_{3j} \in (r_{t_3} \pm \delta_3) && (3.17) \\ &= 0 \text{ otherwise} && (3.18) \end{aligned}$$

and

$$\begin{aligned} Pr(z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) &= \frac{1}{2\delta_4} \text{ if } t^{Ad} = z_j^{Ad}, r^*_{4j} \in (r_{t_4} \pm \delta_4) && (3.19) \\ &= 0 \text{ otherwise.} && (3.20) \end{aligned}$$

Since rank swapping is done independently to obtain values z_3 and z_4, we can simply multiply the terms in Equations 3.18 and 3.20 to obtain the joint probability

$$\begin{aligned} Pr(z_{3j}, z_{4j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) &= \frac{1}{\delta_3}\frac{1}{\delta_4} \text{ if } t^{Ad} = z_j^{Ad}, \; (r_{t_3} - \delta_3, r_{t_3} + \delta_3), \text{ and } (r_{t_4} - \delta_4, r_{t_4} && (3.21) \\ &= 0 \text{ otherwise.} && (3.22) \end{aligned}$$

Combining the components for z_{1j}, z_{2j}, z_{3j}, and z_{4j} we arrive at the following expression for C_{SDL} :

$$\begin{aligned} C_{SDL} &= Pr(z^{Ap}|J = j, t, Z^{Ad}) && (3.23) \\ &= \phi_{1j,\tau_{1j}}\phi_{2j,\tau_{2j}}\frac{1}{\delta_3}\frac{1}{\delta_4} \text{ if } t^{Ad} = Z^{Ad}, r^*_{3j} \in (r_{t_3} \pm \delta_3), r^*_{4j} \in (r_{t_4} \pm \delta_4) && (3.24) \\ &= 0 \text{ otherwise} && (3.25) \end{aligned}$$

We can use the ideas presented above to formulate components corresponding to additional available perturbed variables, Z^{Ap}, when conditional quantile regression predictions and hot deck imputation with rank swapping are used to generate values in synthetic records. When other SDL methods are used to generate synthetic data, it seems feasible to extend the ideas presented here and in Reiter (2005) to formulate components of C_{SDL}. In particular, it should be straight forward to extend the formulation of $Pr(z_{1j}|J = j, t, Z^{Ad})$ and $Pr(z_{2j}|z_{1j}, J = j, t, Z^{Ad})$ in Equations 3.14 and 3.15 to synthetic values that are generated using predictions from any conditional model, provided distibutional properties of model estimates and subsequent predictions are known or can be derived. Reiter (2005) presents possible formulations of $Pr(z_j^{Ap}|J = j, t, Z^{Ad})$ when swapping, re-categorizing, and noise addition are used.

B: $B_{SDL} = Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad})$

Variables z_j^U are variables that are unavailable before release. They include z_{5j}^{Up} which is perturbed using quantile regression predictions before release, z_{6j}^{Up} which is perturbed using hot deck and rank swapping before release, and z_{7j}^{Ud} w hich is unperturbed before release. Recall, an SDL intruder is assumed to know how values are generated for each variable in Z. As for C_{SDL}, we assume the intruder uses this knowledge to formulate B_{SDL}. This implies the intruder's joint conditional probability of observing z_5, z_6, z_7, will be formulated using information about the statistical disclosure limitation procedure. In our hypothetical data set, Z^U is comprised of z_5, z_6, z_7, the variables with values unknown to the intruder prior to data release. We rewrite the joint distribution of z_5, z_6, z_7, as a series of conditional distributions. For every $j = 1, ..., r$,

$$\begin{aligned} B_{SDL} &= Pr(z_j^U|z_j^{Ap}, J = j, t, Z^{Ad}) \\ &= Pr(z_{5j}, z_{6j}, z_{7j}|z_j^{Ap}, J = j, t, Z^{Ad}) \\ &= Pr(z_{5j}|z_{1j}, z_{2j}, J = j, t, Z^{Ad}) \\ &\quad \times Pr(z_{6j}|z_{1j}, z_{2j}, J = j, t, Z^{Ad}) \\ &\quad \times Pr(z_{7j}|J = j, t, Z^{Ad}). \end{aligned} \tag{3.26}$$

In the proposed SDL procedure, z_5 is generated using quantile regression predictions conditional on z_{1j}, z_{2j}, and Z^{Ad}, z_6 is generated using hot deck and rank swapping conditional on z_{1j} and z_{2j}, and z_7 is left unperturbed in the released data. Based on this, we divide the variables in U into perturbed and unperturbed just as for available variables, i.e. $z_5, z_6 \in Up$ and $z_7 \in Ud$. In the following paragraphs, we first consider $Pr(z_{5j}|z_j^{Ap}, J = j, t, Z^{Ad})$, then $Pr(z_{6j}|z_j^{Ap}, J = j, t, Z^{Ad})$, and finally $Pr(z_{7j}|J = j, t, Z^{Ad})$.

B1: z_{5j}^{Up}

Variable z_{5j}^{Up} is unavailable to the intruder before release and is perturbed using quantile regression predictions before release. We consider formulating the conditional probability of observing z_5 in a similar fashion as the conditional probability corresponding to z_1. Unlike the probability of observing z_1, the intruder does not have information on z_5 (or any variables in U) prior to data release, i.e. s/he does not have values t_5, t_6, or t_7. Therefore, $Pr(z_{5j}|z_j^{Ap}, J = j, t, Z^{Ad}) = \phi_{5j,\tau_{5j}}$ cannot be evaluated as in Equation 3.14 using z_{5j} and t_5.

The intruder does know that z_{5j} is generated using:

$$z_{5j} = \begin{pmatrix} Z_j^{Ad} \\ z_{1j} \\ z_{2j} \end{pmatrix} \hat{\beta}_5(\tau_{5j}),$$

for every $j = 1, ..., r$, but does not know the value $\hat{\beta}_5(\tau_{5j})$. This parameter estimate could be estimated by the intruder by fitting the corresponding quatile regression model using values in Z. The estimate of $\hat{\beta}_5(\tau_{5j})$ could then be used to compute a predicted value for the target on this variable, $\hat{t}_5$, based on values of t_1, t_2, and t^{Ad}.

The SDL intruder's predicted $\hat{t}_5$ is not the exact value released by the agency on the target's record (due to estimated quantile regression estimates). How similar or different these values are will depend on the accuracy of the intruder's estimate and the value $\hat{\beta}_5(\tau_{5j})$ estimated by the agency using the original data set. In other words, if the relationship between z_5 and z_1, z_2, and Z^{Ad} are preserved very accurately in the released data, this would result in accurate estimated coefficients for the intruder, and an accurate predicted value of the target's released value. The intruder can either act as if $\hat{t}_5$ is equal to the target's released value or account for additional error introduced by estimating the model coefficients using Z rather than Y. We consider the former scenario, but recognize resulting probability estimates may differ when considering the latter.

Using the estimated value of $\hat{t}_5$, the intruder can choose to act as if this estimate is the target's value of t_5 and procede as above for z_1 and z_2. Plugging $\hat{t}_5$ in for t_5 to evaluate the Normal density, we obtain the following expression for the probability of observing z_{5j} for every $j = 1, ..., r$:

$$Pr(z_{5j}|z_{2j}, z_{1j}, J = j, t, Z^{Ad}) = \phi_{5j,\tau_{5j}}$$

$$= \begin{cases} \phi\left(\frac{z_{5j}-t_5}{\omega(\tau_{5j})\sqrt{\frac{1}{n}(Z_j^{Ad}\ z_{1j}\ z_{2j})(Z_j^{Ad}\ z_{1j}\ z_{2j})^T}}\right), & t^{Ad} = z_j^{Ad} \\ 0 & \text{otherwise} \end{cases} \tag{3.27}$$

where

$$\tau_{5j} = \begin{cases} \tau_{5j}^* & \text{intruder knows value of randomly drawn } \tau_{5j}^* \text{ used to generate } z_{5j} \\ \hat{\tau}_{5j,intruder}^* & \text{intruder estimates value of randomly drawn } \tau_{5j}^* \text{ used to generate } z_{5j}. \end{cases}$$

Alternatively, the intruder may compare the target values of t^{Ad} and $\hat{t}_5$ to values in Z. Suppose there are n_{t_5} records with $t^{Ad} = Z^{Ad}$ and $z_5 = \hat{t}_5 \pm \gamma_5$, some $\gamma_5 > 0$, then the probability can be formulated as

$$Pr(z_{5j}|z_j^{Ap}, J = j, t, Z^{Ad}) = \begin{cases} \frac{1}{n_{t_5}}, & t^{Ad} = Z^{Ad} \text{ and } z_5 = \hat{t}_5 \pm \gamma_5 \\ 0, & \text{otherwise.} \end{cases} \tag{3.28}$$

The size of γ_k depends on the amount of error the intruder attributes to estimating the regression coefficient estimates. The intruder would likely be more willing to act more certain about $\hat{t}_5$ as an estimate of t_5 if the model parameters are well estimated using the released data set. However, the intruder might want to make computation reasonable and choose γ_k so that n_{t_5} is not too large or too small. Namely, if there are few observations close to the predicted target value, then a value of γ_k that is somewhat large would ensure n_{t_5} is not too small. In particular, one would not want to take the chance of eliminating potential matches that could be the target through a choice of n_{t_5} that is too small.

B2: z_{6j}^{Up}

Variable z_{6j}^{Up} is unavailable to the intruder before release and is perturbed using hot deck and rank swapping before release. Recall, hot deck and rank swapping are combined to generate values for z_6 in the released data. We consider formulating $Pr(z_{6j}|z_{1j}, z_{2j}, J = j, t, Z^{Ad})$ in a similar manner as the corresponding probability statements for z_3 and z_4. In the case of z_6, however, we do not have the target value t_6 to use to compute the rank of this variable in the target's record. To use the previous formulation, the intruder would need to estimate the rank of t_6. This could be done by estimating the value of t_6, according to some model, then computing its rank relative to values of z_6 in the relased data, or perhaps by modeling

the ranks themselves, r^*_{6j}, conditional on other variables in Z. Investigating the best way to estimate the rank of t_6 in the target's record is left to outside research. In a simulation study, we consider estimating t_6 based on a model and computing the rank of the estimated value with respect to values of z_6.

Regardless of how this is done, if the intruder obtains an estimate of the rank of t_6, $\hat{r}_{t_6}$ say, then s/he can use this value to evaluate

$$Pr(z_{6j}|z_{1j}, z_{2j}, J = j, t, Z^{Ad}) = \begin{cases} \frac{1}{2\delta_6}, & t^{Ad} = Z^{Ad}, r^*_{6j} \in (\hat{r}_{t_6} \pm \delta_6) \\ 0, & \text{otherwise} \end{cases} \tag{3.29}$$

for some $\delta_6 > 0$.

B3: z_{7j}

Variable z_{7j} is unavailable to the intruder before release and is unperturbed before release. To compute $Pr(z_{7j}|J = j, t, Z^{Ad})$, we rely on an argument presented in Reiter (2005). The author presents the conditional probability as an integral of the joint probability of z^U_j and y^U_j over values of y^U_j as follows:

$$Pr(z^U_j|J = j, t, Z^{Ad}) = \int Pr(z^U_j|y^U_j, J = j, t, Z^{Ad}) Pr(y^U_j|J = j, t, Z^{Ad}) dy^U_j.$$

The author points out that if variables in U remain unperturbed, i.e. $U = Ud$, then $Pr(z^U_j|y^U_j, J = j, t, Z^{Ad}) = 1$, so the entire integral integrates to 1. For our purposes, we set $Pr(z_{7j}|J = j, t, Z^{Ad}) = 1$, for all $j = 1, ..., r$, assuming the intruder knows values on z_7 are all left unperturbed from their original values.

A: $A_{SDL} = Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad})$

Variables z^C_i are variables the intruder cannot know whit certainty after release. Variables in z^{Ap}_i and z^U_i are in z^C_i, so these variables include $z^{Ap}_{1i}, z^{Ap}_{2i}, z^{Ap}_{3i}, z^{Ap}_{4i}, z^{Up}_{5i}, z^{Up}_{6i}$, and z^{Ud}_{7i}. This component computes the probability associated with all records except the target's. In Section 3.2.2, we introduce A_{SDL} as the third term in the right hand side of Equation 3.8, where

$$A_{SDL} = Pr(z^C_1, ..., z^C_{j-1}, z^C_{j+1}, ..., z^C_r|z^C_j, J = j, t, Z^{Ad}).$$

Assuming records are independent, this expression simplifies to the product of conditional probabilities (Reiter 2005):

$$Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad}) = \prod_{\substack{i=1,...,r \\ i \neq j}} Pr(z_i^C | z_j^C, J = j, t, Z^{Ad}) \qquad (3.30)$$

Since records are independent, then for $i \neq j$,

$$Pr(z_i^C | z_j^C, J = j, t, Z^{Ad}) = Pr(z_i^C | z_i^{Ad}).$$

Substituting this into Equation 3.30 and rewriting the product, we obtain

$$Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad}) = \frac{\prod_{i=1,...,r} Pr(z_i^C | z_i^{Ad})}{Pr(z_j^C | z_j^{Ad})}.$$

Substituting this into Equation 3.1 for $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad})$ results in further simplifications that occur from summing over all records in the denominator of 3.1. As a result, the above substitution is equivalent to substituting $\frac{1}{Pr(z_j^C | z_j^{Ad})}$ into 3.1 for $Pr(z_1^C, ..., z_{j-1}^C, z_{j+1}^C, ..., z_r^C | z_j^C, J = j, t, Z^{Ad})$.

This implies that only $Pr(z_j^C | z_j^{Ad})$ is needed to compute A_{SDL}. This probability is decomposed into conditional probabilities according to available perturbed, unavailable perturbed, and unavailable unperturbed variables as in the preceding sections.

$$\begin{aligned} Pr(z_j^C | z_j^{Ad}) &= Pr(z_j^{Ap}, z_j^{Up}, z_j^{Ud} | z_j^{Ad}) \\ &= Pr(z_j^{Ap} | z_j^{Ad}) Pr(z_j^{Up} | z_j^{Ap}, z_j^{Ad}) Pr(z_j^{Ud} | z_j^{Ap}, z_j^{Up}, z_j^{Ad}). \end{aligned}$$

Recall that for unavailable, unperturbed variables z_j^{Ud}, the corresponding probability is set to 1, resulting in further simplification of $Pr(z_j^C | z_j^{Ad})$ to

$$Pr(z_j^C | z_j^{Ad}) = Pr(z_j^{Ap} | z_j^{Ad}) Pr(z_j^{Up} | z_j^{Ap}, z_j^{Ad}). \qquad (3.31)$$

Under the SDL method of this section, the probability becomes

$$\begin{aligned} Pr(z_j^C | z_j^{Ad}) &= Pr(z_{1j}^{Ap} | z_j^{Ad}) Pr(z_{2j}^{Ap} | z_{1j}^{Ap}, z_j^{Ad}) Pr(z_{3j}^{Ap}, z_{4j}^{Ap} | z_{2j}^{Ap}, z_{1j}^{Ap}, z_j^{Ad}) \\ &\quad \times Pr(z_{5j}^{Up} | z_j^{Ap}, z_j^{Ad}) Pr(z_{6j}^{Up} | z_{5j}^{Up}, z_j^{Ap}, z_j^{Ad}). \end{aligned}$$

To evaluate this probability, consider the same SDL methods as those used as in the previous sections. The probabilities in Equation 3.31 are no longer conditioned on target information, t, or $J = j$. Therefore, instead of using values from the target information t, values in each record are used. For $j = 1, ..., r$ the probabilities are listed here, followed by a brief discussion:

$$
\begin{aligned}
z_j^{Ap}: \qquad Pr(z_{1j}|z_j^{Ad}) = \;& \tilde{\phi}_{1j,\tau_{1j}} = \phi\left(\frac{z_{1j}-\hat{z}_{1j,\tau}}{\omega(\tau_{2j})\sqrt{\frac{1}{n}Z_j^{Ad^T}Z_j^{Ad}}}\right) \\
Pr(z_{2j}|z_{1j}, z_j^{Ad}) = \;& \tilde{\phi}_{2j,\tau_{2j}} = \phi\left(\frac{z_{2j}-\hat{z}_{2j,\tau_{2j}}}{\omega(\tau_{2j})\sqrt{\frac{1}{n}(Z_j^{Ad}\;\; z_{1j})(Z_j^{Ad}\;\; z_{1j})^T}}\right) \\
Pr(z_{3j}|z_{2j}, z_{1j}, z_j^{Ad}) = \;& \tfrac{1}{2\delta_3}, \quad \hat{r}_{3j} \in (r_{3j}-\delta_3, r_{3j}+\delta_3) \\
Pr(z_{4j}|z_{2j}, z_{1j}, z_j^{Ad}) = \;& \tfrac{1}{2\delta_4}, \quad \hat{r}_{4j} \in (r_{4j}-\delta_4, r_{4j}+\delta_4) \\
z_j^{Up}: \quad Pr(z_{5j}|z_{2j}, z_{1j}, z_j^{Ad}) = \;& \tilde{\phi}_{5j,\tau_{5j}} = \tilde{\phi}\left(\frac{z_{5j}-\hat{z}_{5j,\tau_{5j}}}{\omega(\tau_{5j})\sqrt{\frac{1}{n}(Z_j^{Ad}\;\; z_{1j}\;\; z_{2j})(Z_j^{Ad}\;\; z_{1j}\;\; z_{2j})^T}}\right) \\
Pr(z_{6j}|z_{2j}, z_{1j}, z_j^{Ad}) = \;& \tfrac{1}{2\delta_6}, \quad \hat{r}_{6j} \in (r_{6j}-\delta_6, r_{6j}+\delta_6)
\end{aligned}
\tag{3.32}
$$

The formulations of the components of A_{SDL} are quite similar to the components for B_{SDL} and C_{SDL} since the SDL method and intruder's knowledge are the same. They differ in due to conditioning only on observed values in record j, rather than on the target's information and on $J = j$, the j^{th} record belonging to the target. The terms $\hat{z}_{kj,\tau_{kj}}$ are defined as before, as the quantile regression predictions for variable k at the τ_{kj} quantile, computed using released values and estimated parameter estimates obtained from the released data. The probabilities associated with the variables that imputed using hot deck and rank swapping remain at $\frac{1}{2\delta_k}$ when the rank of $\hat{z_{kj}}$, $\hat{r}_{kj}$, falls in the interval $(r_{kj}-\delta_k, r_{kj}+\delta_k)$. The value of δ_k is not likely released by the agency. The intruder trades off taking large values of δ_k to cover the true match and a small value of δ_k that is more computationally feasible. Future work could examine ways to estimate δ_k.

3.2.4 Formulation of components for the NAIVE intruder

The naive intruder is one who knows the values of the target's available variables, but has no other prior knowledge. Only after the data are released does the naive intruder gather posterior information. The intuder is naive both in the sense that s/he only has posterior knowledge of the data and in statistical sophistication to identify the target. This intruder can use the released data and values on t to identify the target with records that have values close to the values in t. This can be done in several ways. Here we consider a simple scenario in which the intruder measures the distance between values in the released data and in t on

available variables. The intruder can choose to identify the target with the closest values or choose from among the closest.

$$\textbf{C: } C_{naive} = Pr(z_j^{Ap} | J = j, t, Z^{Ad})$$

The variables z_j^{Ap} are available perturbed variables that are perturbed using quantile regression prior to release. The intruder can choose to compute component C_{naive} by assigning equal probability to records with values on Z^{Ad}, Z^{Ap} within Euclidean distance η, say, of the target values t^{Ad}, t^{Ap}. If the distance between the target and each record j in the released data set must be less than η_k for each variable $k \in Ap$, then jointly, the probability of observing z_j^{Ap} given the j^{th} record is the target's and information on variables Z^{Ad} can be formulated as:

$$C_{naive,j} = \begin{cases} \prod_{k \in Ap} \frac{1}{n_{t_k}}, & t^{Ad} = z_j^{Ad} \text{ and } z_{kj} \in (t_k - \eta_k, t_k + \eta_k) \\ 0, & \text{otherwise,} \end{cases} \tag{3.33}$$

where n_{t_k} =number of records with $z_{kj} \in (t_k - \eta_k, t_k + \eta_k)$. Other distance measures and other scenarios will lead to different formulations of C_{naive}. Comparing accuracy of the probability of identification, and hence the effect that any such formulation would have on disclosure risk among formulations can be done.

$$\textbf{B: } B_{naive} = Pr(z_j^{U} | z_j^{Ap}, J = j, t, Z^{Ad})$$

Variables z_j^U are unavailable variables. They include both perturbed and unperturbed variables unavailable to the intruder before release. If the intruder uses the method described for C_{naive}, it seems reasonable this method would be used for unperturbed variables to compute B_{naive} as well. However, the intruder does not posess the target's values for t_5, t_6, or t_7.

The intruder can consider the associations between variables Z^U and Z^A in the released data set to determine which records are more likely to belong to the target, among records with values $z_{kj}^{Ap} \in (t_k - \eta_k, t_k + \eta_k)$, $k = 1, 2, 3, 4$. The intruder can also attempt to characterize the conditional distribution of Z^U given $Z^A, J = j$, and t using regression analysis and inference to compute the conditional probability of observing z_j^U in record j. The intruder might choose a plausible standard distribution s/he believes the values in Z^U to belong to and compute the probability of observing values z_{kj}, though it could become rather complicated to incorporate conditioning information. For simplicity, we set B_{naive} equal to 1 for records with $z_j^{Ad} = t^{Ad}$ and 0 otherwise.

$$\textbf{A: } A_{naive} = \frac{1}{Pr(z_j^C | z_j^{Ad})}$$

Variables z_j^C include those the intruder cannot know with certainty after the data are released. This probability is based on all records in the released data set except the target's record. Again, the naive intruder has knowledge of variables Z^C only after the data are released. Recall that for $A_{naive} = \frac{1}{Pr(z_j^C | z_j^{Ad})}$, the probability is not conditional on the SDL method or the assumption that the j^{th} record belongs to the target. In this context, the SDL intruder can estimate the distribution that s/he believes the observations have on variables Z^C, assume observations belong to a well known distribution, or use some other empirical method to assess the probability. In our application to Public Use Microdata Sample data in Section 4.3, we suppose that the intruder computes this probability based on variables z_k^C belonging to a normal distribution. A_{naive} can also be set to 1 if we assume the intruder will not make any distributional assumptions about variables in the released data set.

3.2.5 Formulation of components for the AVERAGE intruder

An *average* intruder is one with some knowledge about the original data prior to its release and the target's values on available variables. This intruder does not have specific knowledge of the SDL methods used to generate the synthetic data set. Posterior information is gained after the synthetic data set is released. Prior knowledge is combined with posterior information to assist the average intruder's attempt to identify a target in the released data set.

In order to assess the threat from an average intruder, the agency must consider what information is available to the intruder prior to data release. Prior information could contain specific variable values, including possible information on all of the variables on which the agency intends to release synthetic data. It could also include more general information such as bounds for the target on certain variables, or membership to a broader category. The possibilities are too vast for us to consider exhaustively. In an application to a Public Use Microdata Set in Section 4.3, we consider a scenario in which results from a regression analysis are available to the intruder prior to data release. An agency might publish regression estimates in a report about the data base. A journal article or other report might describe associations between variables or the result of a study. The measures of disclosure risk in this section are based on this scenario. It would be useful for the agency to consider other available prior information to fully assess disclosure risk, though we hope to cover a wide range of scenarios by computing disclosure risk for an SDL intruder and a naive intruder in

addition to the average intruder.

In this section, formulations for the components of disclosure risk, C_{avg}, B_{avg}, and A_{avg} are presented.

$$\textbf{C: } C_{avg} = Pr(z_j^{Ap} | J = j, t, Z^{Ad})$$

The variables z_j^{Ap} are variables that are available to the intruder prior to release and perturbed before release. Recall that the values of available perturbed variables are compared with the target's values on variables in Ap to formulate the conditional probability $C_{avg} = Pr(z_j^{Ap} | J = j, t, Z^{Ad})$. Recall the scenario presented in Section 3.2.2. The available perturbed variables we consider are $y_1, ..., y_4$ and the corresponding synthetic values $z_1, ..., z_4$. In this section, we consider prior information to include parameter estimates from these linear models:

$$\begin{aligned} Y_{1j} &= Y_j^{Ad}\beta_1 + \epsilon_1, \\ Y_{2j} &= Y_j^{Ad}\beta_2 + \epsilon_2, \\ Y_{3j} &= Y_j^{Ad}\beta_3 + \epsilon_3, \text{ and} \\ Y_{4j} &= Y_j^{Ad}\beta_4 + \epsilon_4. \end{aligned} \tag{3.34}$$

The released estimates, $\hat{\beta}_k$, $k = 1, 2, 3, 4$, are computed by the agency using the original data to fit these models. Assuming the errors are independent and identically distributed normal random variables with zero mean and variance σ_k^2, models imply that given Y^{Ad}, each y_k is normally distributed with mean $Y^{Ad}\beta_k$ and variance σ_k^2, or $y_k \sim N(Y^{Ad}\beta_k, \sigma_k^2)$.

Since the synthetic data set is produced with the goal that values on Z_k reproduce marginal and conditional distributions of values on variable Y_k, it is reasonable to extend the distributional assumptions from y_k to z_k, keeping in mind extra variability may exist in the values z_k. Namely, normality can be extended to assume $z_k \sim N(Z^{AD}\beta_k, \sigma_k^2 d)$. Notice the variance of y_k, σ_k^2, is multiplied by a constant d. The multiplier d can be thought of as an SDL correction factor used to account for the variability in z_k that results from the SDL method. It implies that the variance of z_k, is $\sigma_{z_k}^2 = \sigma_k^2 d$.

The SDL correction factor d is necessarily positive. If it is larger than 1, the variance of values on z_k is increased from the variance of values y_k. This implies that the SDL procedure introduced additional variability to the synthetic values of variable k. If d is equal to 1,

this implies there is no change in the variance due to the SDL method. If d is between 0 and 1, this implies the SDL method had the effect of reducing the variability. This could happen if the SDL method were to use predicted values from a mean regression model to generate values for synthetic data. Generated values would approximate the mean at any given predictor value, likely decreasing the variability between generated values.

The value of d may be released by the agency as a measure of data utility, allowing users to include the correction factor in their analyses. If the value of d is not released, but the estimate $\hat{\sigma}_k^2$, calculated from y_k, is released, then d can be deduced by calculating the variance estimate in the released values to obtain $\hat{\sigma}_{z_k}^2$ and using both variance estimates to solve $\hat{\sigma}_{z_k}^2 = \hat{\sigma}_k^2 d$ to obtain an estimate $\hat{d}$. If neither $\hat{\sigma}_k^2$ or d is released, the variance estimate from synthetic data on the released variables, $\sigma_{z_k}^2$, can be used, keeping in mind that it differs by an unknown factor d from σ_k^2. The value of d may be set to some constant based on SDL literature discussing data utility in data sets generated using various SDL methods.

Initially, suppose that the true values of $\beta_k,\ \sigma_k^2$, and d are known and released by the agency. The average intruder uses this knowledge to compute each portion of A_{avg}. Because this probability is conditional on the j^{th} record belonging to the target, the distributional assumption for z_{kj}^{Ap} becomes $z_{kj}^{Ap} \sim N(t^{Ad}\beta_k, \sigma_k^2 d)$, since $t_k \sim N(t^{Ad}\beta_k, \sigma_k^2 d)$ is implied by the linear model assumptions. This leads to formulating the conditional probability as:

$$Pr(z_{kj}^{Ap}|J=j,t,Z^{Ad}) = \begin{cases} \phi\left(\frac{z_{kj}^{Ap} - t_k^{Ad}\beta_k}{\sigma_k\sqrt{d}}\right), & t^{Ad} = z_j^{Ad} \\ 0, & \text{otherwise.} \end{cases} \tag{3.35}$$

Using this formulation, the average intruder's prior knowledge is incorporated through values of $\beta_k,\ \sigma_k$, and d, while posterior knowledge is incorporated through the values z_{kj}^{Ap}.

Since it is unlikely that the agency has released the true parameter values β_k or σ_k, it is more likely that estimates $\hat{\beta}_k$ and $\hat{\sigma}_k$ will be released. If an estimate of d, $\hat{d}$, is also released, the estimated values can be substituted in to Equation 3.35 to obtain:

$$Pr(z_{kj}^{Ap}|J=j,t,Z^{Ad}) = \begin{cases} \phi\left(\frac{z_{kj}^{Ap} - t_k^{Ad}\hat{\beta}_k}{\hat{\sigma}_k\sqrt{\hat{d}}}\right), & t^{Ad} = z_j^{Ad} \\ 0, & \text{otherwise.} \end{cases} \tag{3.36}$$

The more statistically savvy intruder, however, can instead use inference on the estimated

regression coefficients to formulate the probability as:

$$Pr(z_{kj}^{Ap}|J=j,t,Z^{Ad}) = \begin{cases} \phi\left(\frac{z_{kj}^{Ap}-t^{Ad}\hat{\beta}_k}{\hat{\sigma}_k\sqrt{\frac{\hat{d}}{n}t^{Ad}t^{Ad^T}}}\right), & t^{Ad}=z_j^{Ad} \\ 0, & \text{otherwise.} \end{cases} \tag{3.37}$$

In both Equations 3.36 and 3.37, prior knowledge in the form of parameter estimates are combined with posterior knowledge of the values z_{kj}^{Ap} to compute the conditional probability.

Using the same rationale for variables z_2, z_3, and z_4, and assuming independence between the variables in Ap, we formulate C_{avg} as:

$$C_{avg} = Pr(z_{kj}^{Ap}|J=j,t,Z^{Ad}) = \begin{cases} \prod_{k\in Ap}\phi_{kj}, & t^{Ad}=z_j^{Ad} \\ 0, & \text{otherwise,} \end{cases} \tag{3.38}$$

where $\phi_{kj} = \phi\left(\frac{z_{kj}^{Ap}-t_k^{Ad}\hat{\beta}_k}{\hat{\sigma}_k\sqrt{\frac{\hat{d}}{n}t^{Ad}t^{Ad^T}}}\right)$ and $\hat{\sigma}^2_{z_k}$ is substituted for $\hat{\sigma}_k^2\hat{d}$. Alternatively, we could model z_1, $z_2|z_1$, $z_3|z_2,z_1$, and $z_4|z_3,z_2,z_1$.

Note that many other possible scenarios of prior information can be used to compute the disclosure risk for the average intruder. This one is used to illustrate formulation based on released regression estimates. It is also possible that the same prior information can be used in different ways to formulate A_{avg}.

B: $B_{avg} = Pr(z_j^U|z_j^{Ap}, J=j,t,Z^{Ad})$

Variables z_j^U are unavailable to the intruder before data are released. These variables include perturbed and unperturbed variables. Recall that the unavailable variables in U include y_5^{Up}, y_6^{Up}, and y_7^{Ud}, and corresponding synthetic values z_5^{Up}, z_6^{Up}, and z_7^{Ud}. We formulate an expression for B_{avg} supposing that, in addition to the models for Y^{Ap}, the intruder has prior information about the regression coefficients and variance of values in these models:

$$\begin{aligned} Y_{5j} &= (Y^{Ad}\ Y^{Ap})\beta_5 + \epsilon_5, \\ Y_{6j} &= (Y^{Ad}\ Y^{Ap})\beta_6 + \epsilon_6, \text{ and} \\ Y_{7j} &= (Y^{Ad}\ Y^{Ap})\beta_7 + \epsilon_7. \end{aligned} \tag{3.39}$$

For ease of notation, we write $Y^A = (Y^{Ad}, Y^{Ap})$ for the remainder of this section. Since the models are assumed to represent relationships in the data accurately, y_k^{Up} is assumed to

be normally distributed such that $y_k^{Up} \sim N(Y^A\beta_k, \sigma_k^2)$. This implies that $t_k^{Up} \sim N(t^A\beta_k, \sigma_k^2)$ and leads to $z_k^{Up} \sim N(Y^A\beta_k, \sigma_k^2)$ through similar reasoning as presented for the formulation of C_{avg}.

In the case of the SDL and naive intruders, formulating an expression for component B could not follow directly from the expressions for component C because the target values on unavailable perturbed variables t_5 and t_6 are unavailable and unknown to the intruder. In the scenario here, however, values t_5 and t_6 are not necessary when prior information is from the linear regression models. If the intruder possesses estimates $\hat{\beta}_k, \hat{\sigma}_k$, and $\hat{d}$, the conditional probability of observing z_{kj} can be formulated as:

$$Pr(z_{kj}^{Up}|J = j, t, Z^{Ad}) = \phi\left(\frac{z_k^{Up} - t^A\hat{\beta}_k}{\hat{\sigma_k}\sqrt{\frac{\hat{d}}{n}t^A t^{A^T}}}\right). \tag{3.40}$$

Recall that values of variables in Ud remain unperturbed, so the conditional probability of observing z_7^{Ud} is set to 1. Therefore, using the expression in **??** and assuming independence between z_5 and z_6, B_{avg} can be formulated as:

$$B_{avg} = Pr(z_j^U|J = j, t, Z^{Ad}) = \begin{cases} \prod_{k\in Up}\phi_{kj}, & z_{kj}^{Ad} = t^{Ad} \\ 0, & \text{otherwise,} \end{cases} \tag{3.41}$$

where $\phi_{kj} = \phi\left(\frac{z_{kj}^{Up} - t^A\hat{\beta}_k}{\hat{\sigma_k}\sqrt{\frac{\hat{d}}{n}t^A t^{A^T}}}\right)$ and $\hat{\sigma}_{z_k}^2$ is substituted for $\hat{\sigma}_k^2\hat{d}$.

$$\textbf{A: } A_{avg} = Pr(z_1^C, , z_{j-1}^C, z_{j+}^C, , z_r^C|z_j^C, J = j, t, Z^{Ad})$$

Variables in this component include available perturbed, unavailable perturbed, and unavailable unperturbed variables on all records except the intruder's. Recall from Section 3.2.3 that A_{avg} simplifies to $\frac{1}{Pr(z_j^C|z_j^{Ad})}$ when plugged into Equation 3.1. This term is not conditional on $J = j$ or t, hence the component A_{avg} corresponds to the probability of observing z_{kj}^C given z_{kj}^{Ad} only. Since the distributional assumptions $z_k^{Ap} \sim N(Z^{Ap}\beta_k, \sigma_k^2 d)$ and $z_l^{Up} \sim N(Z^A\beta_l, \sigma_l^2 d)$ are implied by the linear models in Equation 3.34 and **??**, we use them to compute A_{avg} as follows:

$$A_{avg} = \prod_{k\in Ap}\frac{1}{\phi_{kj}}\prod_{l\in Up}\frac{1}{\phi_{lj}}, \tag{3.42}$$

where $\phi_{kj} = \phi\left(\frac{z_k^{Ap} - \hat{z}_k^{Ap}}{\hat{\sigma}_k\sqrt{\frac{\hat{d}}{n} z_j^{Ad^T} z_j^{Ad}}}\right)$, $\phi_{lj} = \phi\left(\frac{z_l^{Up} - \hat{z}_l^{Up}}{\hat{\sigma}_l\sqrt{\frac{\hat{d}}{n} z_j^{A^T} z_j^{A}}}\right)$, and $\hat{\sigma}^2_{z_k}$ is substituted for $\hat{\sigma}^2_k\hat{d}$.

3.2.6 Summary

Under the framework for computing disclosure risk presented in Duncan and Lambert (1986, 1989) and Reiter (2005), components of disclosure risk were formulated based on various levels on intruder knowledge and decisions. A summary is presented in this section. Results from implementing these measures of disclosure risk are presented in an application to a Public Use Microdata Sample from the Census Bureau in Section 4.3.

The framework for measuring disclosure assumes that an intruder computes the probability of identifying a target in the released data set, which is expressed as $Pr(J = j|t, Z^{Ad})$. Disclosure risk is equated with the probability of identification. If the agency can control the probability of identification to be low for a target, then the disclosure risk is also low. Alternatively, if the probability of identication is the same accross a large number of records, this may prevent the intruder from identifying any record as the target's, resulting in low disclosure risk as well.

To assess disclosure risk, we consider the various types of information or knowledge an intruder has before data release, information gained after release, and various decisions the intruder can make about how to calculate the probability of identification. Such decisions are based on the level of information s/he possesses. Refer to Table 3.3 in which the framework under w hich disclosure risk is formulated based on knowledge and decisions an intruder can make to identify a target record. Disclosure risk is divided into extreme cases based on an *SDL* intruder and a *naive* intruder. An *average* intruder is also considered to give insight into a possibly more common type of intruder. By computing disclosure risk for each type of intruders, we hope to cover a wide range of possibilities, enabling the agency to evaluate risk in a worst case scenario, a best case scenario, and a more common scenario. We have also included something like a best best case scenario using the *super naive* intruder, who bases the probability of identification only on the number of records with matching available unperturbed variables.

Disclosure risk can also be computed for the intruder that makes decisions to compute the components in a simpler manner than s/he has information for. For example, an intruder with accurate and detailed information about the SDL method used can choose to compute

C_{SDL}, B_{SDL}, and A_{SDL} to obtain the probability of identification. Alternatively, an SDL intruder can choose to compute C_{SDL}, but use B_{avg} and A_{avg}, or even set these components to 1, to compute the probability of identification. The average intruder has options too. S/he can compute all components using the average formulations (C_{avg}, B_{avg}, and A_{avg}), or can compute any of these components at the naive level or set any of them equal to 1. The naive intruder can only choose to compute C_{naive}, B_{naive}, and A_{naive} or set these components equal to 1. The naive intruder, however, cannot choose to increase the amount of prior knowledge used to compute any component. In total, there are 3^3 options of combinations of A, B, and C available to the SDL intruder, or 4^3 options if we include setting any component to 1. There are 2^3 (or 3^3) options for the average intruder, and one option (or 2^3) for the naive intruder.

Intruders with other levels of knowledge exist and can make choices to formulate the probability of identification in a different way than we have. We hope to account for the best and worst case scenarios based on SDL knowledge, average knowledge, and naive knowledge by using the formulations presented in this section.

3.3 Data utility

Synthetic data methods are designed to protect confidential data and simultaneously provide useful microdata to data users. In Section 3.1, disclosure risk was discussed. It is necessary to formulate methods for measuring the disclosure risk under various assumptions about intruder knowledge and behavior. Likewise, to assess data utility, various user behaviors can be assumed. User behaviors reflect the potential uses of the released data set. Uses can include hypothesis testing, linear regression or analysis of variance on a set of variables, the goodness of fit for models used, or other measures of association between variables (Shlomo and Young 2006).

For certain users who frequently request data from an agency for a specific purpose, it might be possible to tailor the SDL method to provide data of high utility for the specific purpose. For synthetic microdata sets to be released for general use, however, this is not likely possible. Therefore, general measures of data utility are considered to asses the synthetic data set's utility.

Winkler (2006) discusses data utility as whether or not released data are "fit for use."

Fitness for use is measured by the extent to which the first two moments agree between the original and synthetic data sets. If moment estimates approximately agree, this implies that regression estimates for both data sets will approximately agree as well.

Shlomo and Young (2006) measure data utility using information loss measures. Their work focuses on assessing the data utility of released tables. The authors have developed a software tool in SAS to assess disclosure risk and information loss associated with summary statistics, bias estimates, hypothesis testing, distance metrics, and many more analyses. To implement measures of utility, assumptions are made about how a user will manage the released data. In terms of tables, assumptions include various methods a user might use to replace values of suppressed cells. To measure the difference between original and released data, distance metrics are used to compare values from the original and released data tables. Values include counts in specific cross-classification cells and cell count variance estimates.

Karr et al. (2006) discuss measuring confidence interval overlap and ellipsoid overlap as two narrow measures of utility that capture the difference or similarities between regression results from the original and synthetic data sets. As a broad measure of utility, the authors discuss using a Kullback-Liebler divergence measure applied to the estimated densities from each data set. Further, the utility measures described in this article provide quantitative measures of utility that can be used in the RU Confidentiality Map, discussed in the opening to this chapter, to balance risk and utility of proposed SDL methods. Several reports available through the NISS website (www.niss.org/dgii/techreports.html) discuss measures of data utility, including the mean squared error of estimates from the orginal and synthetic data sets, for use in the RU Confidentiality Map.

A collection of tools is presented here that can be used to assess broad data utility in a released synthetic set. In Section 3.2.1, a collection of plots is suggested to compare distributions in the original data with distributions in the released data to assess general data utility. In Section 3.2.2, comparisons are made between original and synthetic data using regression analysis. While the measures presented here do provide insight into how the synthetic data compare to the original data overall, there remains much work to be done if a quantitative utility measure is desired.

3.3.1 Visual measures

Data utility can be measured in terms of the marginal and conditional distributions of values on variables in the original and synthetic data sets. Empirical plots of the densities and cumulative distributions provide a good visual tool for the agency to assess how well an SDL method preserves univariate distributions from the original to the synthetic data set. Such plots can also provide insight to where the method fails to preserve distributions, allowing the agency to update the SDL method to perform better in this respect.

To visually compare conditional distributions, we consider comparing conditional distributions using boxplots of continuous variables within categories formed by categorical variables. The cumulative distributions can also be compared within categories defined by discrete variables.

In Chapter 4, plots of empirical ditributions, densities, and box plots are used in applications of the proposed SDL method to three confidential data sets.

3.3.2 Quantitative measures

In addition to the visual diagnostics described in the previous section, comparing point estimates, standard errors, and confidence intervals from the original and synthetic data sets provides a measure of data utility. This can be implemented to assess how well statistics are preserved both marginally and conditionally on variables in the synthetic data set. Another tool to assess utility in conditional relationships is to perform regression analyses and compare regression coefficient estimates, their standard errors and confidence intervals. Goodness of fit measures such as the R^2 values can also be compared between the original and synthetic data sets. If the agency is aware that a particular user will attempt to make a decision based on analyses such as hypothesis testing or likelihood ratio testing to compare estimates or models, the agency can perform both analyses and compare the results.

In Chapter 4, results from a regression analysis are presented and discussed to assess data utility in the synthetic data set generated using the proposed SDL method.

CHAPTER 4. APPLICATIONS

4.1 Iowa Department of Revenue application

The initial motivation for research in the area of disclosure limitation and confidentiality protection was originally brought to Professor Vardeman at Iowa State University by the Tax Research and Program Analysis Section of the Iowa Department of Revenue (IDR). The section's hope was that we could help them to produce a method for generating synthetic or artificial tax return records for use in the Iowa Legislative Services Agency (LSA) tax calculators for determining tax revenue and burden. The end goal has been to provide a tool that will allow IDR to produce a dataset for release to LSA who in turn explore various changes to the Iowa tax code by assessing revenue implications of those potential changes. With a synthetic data set, this can be done without requiring direct IDR involvement with each calculation. A description of the individual income tax reurn data base is presented in Section 4.1.1. Details about the method as applied to individual income tax returns at IDR are presented in Section 4.1.2. Results are described in Section 4.1.3.

4.1.1 Individual income tax return data set

The Iowa Department of Revenue (IDR) collects tax returns annually from all tax payers in Iowa. The individual income tax returns contain data on over 1 million tax payer records. The records contain variables that can be found on individual income tax returns. These include tax payer name, social security number, address, date of birth, tax payer filing status (single, filing jointly, etc.), number of dependents, and other demographic variables describing the tax payer. Income variables include wage income, rent income, retirement income, etc. Adjustments include payments to a retirement account, moving expense deductions, and others. Each variable in the data base corresponds with an entry on the Iowa individual income tax return such as the IA-1040, IA-1040EZ, and others. Data are also borrowed from the federal Internal Revenue Department, so some additional variables from the US-1040 and other forms appear in the Iowa data base. A detailed account of the individual income

tax return data base can be found in Huckett (2006).

Work done with the Iowa Department of Revenue focused on the development of methods for creating a synthetic surrogate of the large individual income tax return database (90 variables, 1.3 million records) for a single year. Data summaries are not sufficient due to the complex nature of the tax rules and the need to be able to experiment with subtle changes in them. Thus, a set of record level data containing all necessary variables derived from state and federal tax forms is necessary.

Quantile regression can be used effectively to estimate quantiles of conditional distributions that can in turn be used to simulate values for age and wages given a collection of demographic and tax variables. Realistically reproducing tails of distributions is of particular concern and requires further investigation and development of methodology. Hot deck imputation can be used to associate values on other variables (both discrete and categorical) with partially constructed synthetic records. Rank swapping is also applied.

4.1.2 Models and procedure

In the application for the Iowa Department of Revenue (IDR), we consider using quantile regression methods to simulate income tax return data by computing a predicted response for randomly selected quantiles of dependent variables age and wages conditional on key demographic predictors corresponding to the number of dependents (dep), state filing status (sfs), and county (county) reported on an individual income tax return. Synthetic values of age are generated conditional on the demographic predictors. We impose the following transformation on wages, since the range of values is considerably large:

$$l.wages = \begin{cases} log(wages), & wages \neq 0 \\ 0, & wages = 0. \end{cases} \tag{4.1}$$

Synthetic values of wages are generated conditional on the demographic predictors and age values using parameter estimates in the following quantile regression models:

$$age_\tau(\tau|x) = \xi(\text{county, sfs, dep, } \hat{\beta}) + \epsilon_{age,\tau}$$

and

$$l.wages_\tau(\tau|x) = \xi(\text{county, sfs, dep, age,}\hat{\beta}) + \epsilon_{l.wages,\tau}.$$

Using quantile regression we estimate parameters in these models for randomly selected quantiles $\tau = \tau^*_{age}$ and $\tau = \tau^*_{l.wages}$. At each value of τ^*_{age}, the predicted value $\widehat{age}_{\tau^*_{age}}$ is computed, given values of the demographic predictors. At each value of $\tau^*_{l.wages}$, the predicted value $\widehat{l.wages}_{\tau^*_{l.wages}}$ is computed, given values of the demographic predictors and values of predicted ages, $\widehat{age}_{\tau^*_{age}}$, on corresponding records. In this application, we randomly selected the quantiles τ^*_{age} and $\tau^*_{l.wages}$ from a uniform distribution over the interval $(0, 1)$.

Practically, it is infeasible to compute quantile regression estimates for unique quantiles drawn for two variables, each with over 1 million records. Therefore, we compute predicted age and wages values at randomly drawn quantiles τ^*_{age} and $\tau^*_{l.wages}$ by interpolating between the predicted responses at quantiles directly above and below τ^*_{age} and $\tau^*_{l.wages}$ from the set $\{0.001, 0.01, ..., 0.99, 0.999\}$. In other words, we actually compute quantiile regression estimates at each quantile in the set $\{0.001, 0.01, ..., 0.99, 0.999\}$, then interpolate between predicted values at these quantiles given the demographic predictors in each record to obtain predicted $\hat{age}_{\tau^*_{age}}$ and $\hat{l.wages}_{\tau^*_{l.wages}}$ values.

Since the distribution of wages is so extremely skew right in the upper tail, in addition to quantile regression techniques, we supplement simulation for wages with sampling from a shifted exponential distribution in the upper tail of the distribution.

At the initial stages of this work with the IDR, we found that performing quantile regression on the entire database of individual income tax returns for any given year to be computationally intensive, if not impossible. At that time, we proposed performing hot deck imputation on remaining sensitive variables as a method to impute other variable values from records with similar values on demographic predictors and synthetic values. We introduced additional perturbation using rank swapping. The result in each synthetic record is original values for key demographic predictors, synthetic values for wages and age via quantile regression, and imputed values for additional variables.

Hot deck imputation is performed using the methods described in Chapter 2. We impute federal tax and deductions amond several other variables. Hot deck matching is done within demographic categories defined by levels of the demographic predictors. We match on wages first and select the 20 closest candidates. Among the 20 closest, we match on both age and wages values and select the closest record with respect to age and wages values as the matching record. Then we perform rank swapping to further perturb federal tax and deductions. We compute the sample rank of federal tax, r, in the matching record and draw

random rank r^* from the Uniform distribution centered at r, or Uniform$(r-20, r+20)$. We impute the value of federal tax that has rank r^* in the original records. Following the same procedure for deductions, we compute the sample rank of deductions, q, in the matching record and draw random rank q^* from the Uniform$(q-20, q+20)$ distribution. Finally, we impute the value of deductions with rank q^* from the original records.

The procedure described above is an application of the proposed method presented in Chapter 2. In this application, we retained the original values on demographic predictors: number of dependents, state filing status, and county. We generated synthetic values for variables age and wages using quantile regression models, and we imputed and perturbed values for variables federal tax and deductions (among others) using hot deck and rank swapping. In Section 4.1.3, we present results from implementing this procedure using individual income tax records for tax payers in the state of Iowa. Depending on agency and department interest, future work could investigate further applications of the method.

An individual income tax return contains up to 100 or more variables. In order to produce a synthetic data set for use by the LSA, values for many more variables need to be generated so that the agency can investigate tax policy changes.

4.1.3 Results

As a preliminary check we assess how well the distributions in the synthetic data approximate the distributions in the original data. For work done at the IDR, we examine plotted empirical cumulative distributions of original and synthetic age and log(wages) values. Plots can be seen in Figures 4.1 and 4.2.

Values on age and log(wages) in the synthetic data match the distribution of values in the original and very precisely. Checks were also performed conditional on values of the independent variables (dependents, filing status, and county). Results are very promising (Huckett 2006). In Figures 4.3 and 4.4, the distribution of values in the original and synthetic records for variables federal tax and deductions are plotted.

We are able to maintain the distributions for several variables using the proposed method. It remains to assess how well we maintain joint distributions in the synthetic set. We take steps to do so in another application at the U.S. Census Bureau. It also remains to generate

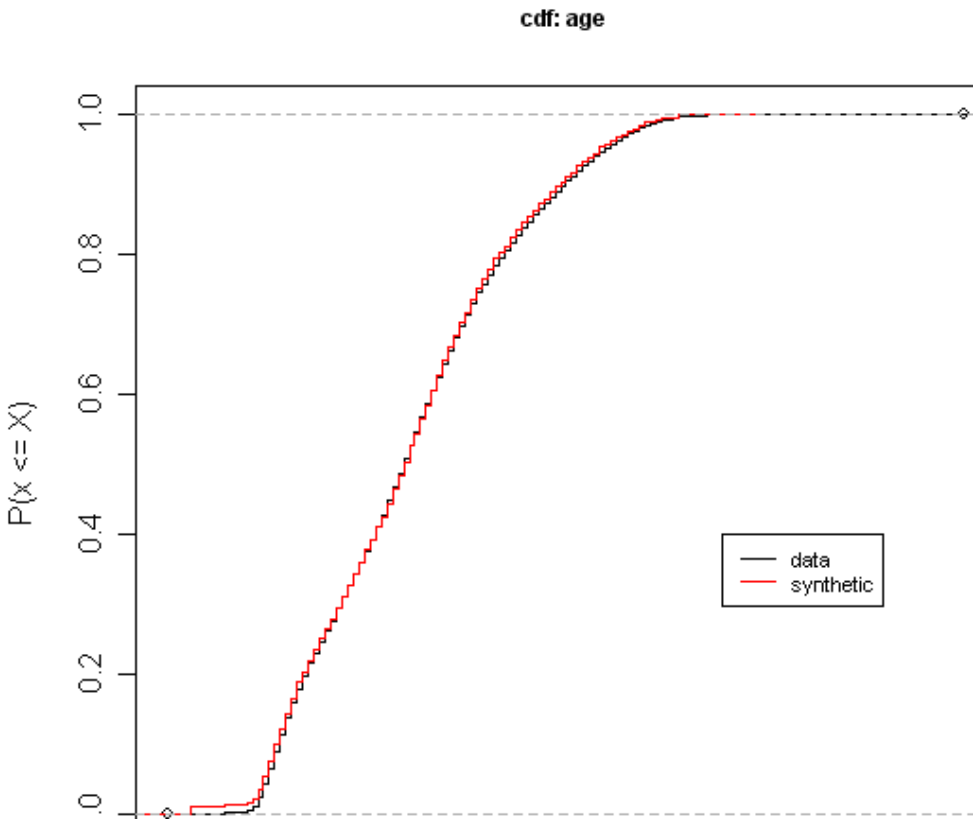

Figure 4.1 Empirical cumulative distribution of age values from original and synthetic Iowa income tax return data.

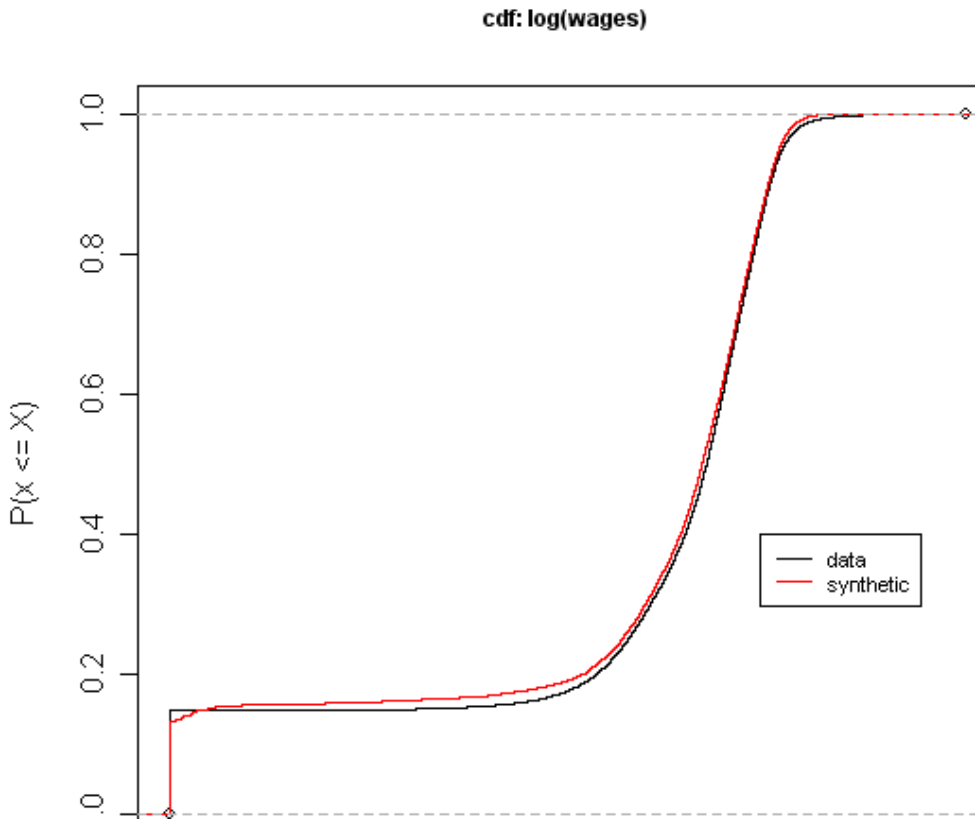

Figure 4.2 Empirical cumulative distribution of log(wages) from original and synthetic Iowa income tax return data.

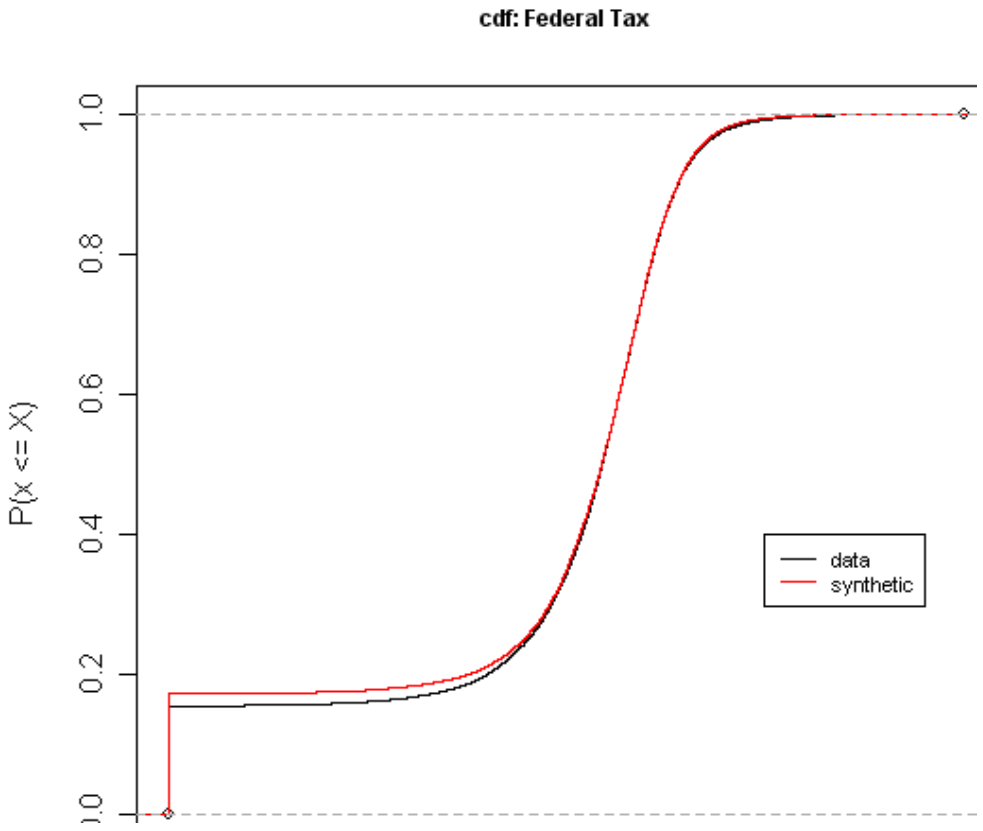

Figure 4.3 Empirical cumulative distribution functions of log(federal tax) from original and synthetic Iowa income tax return data.

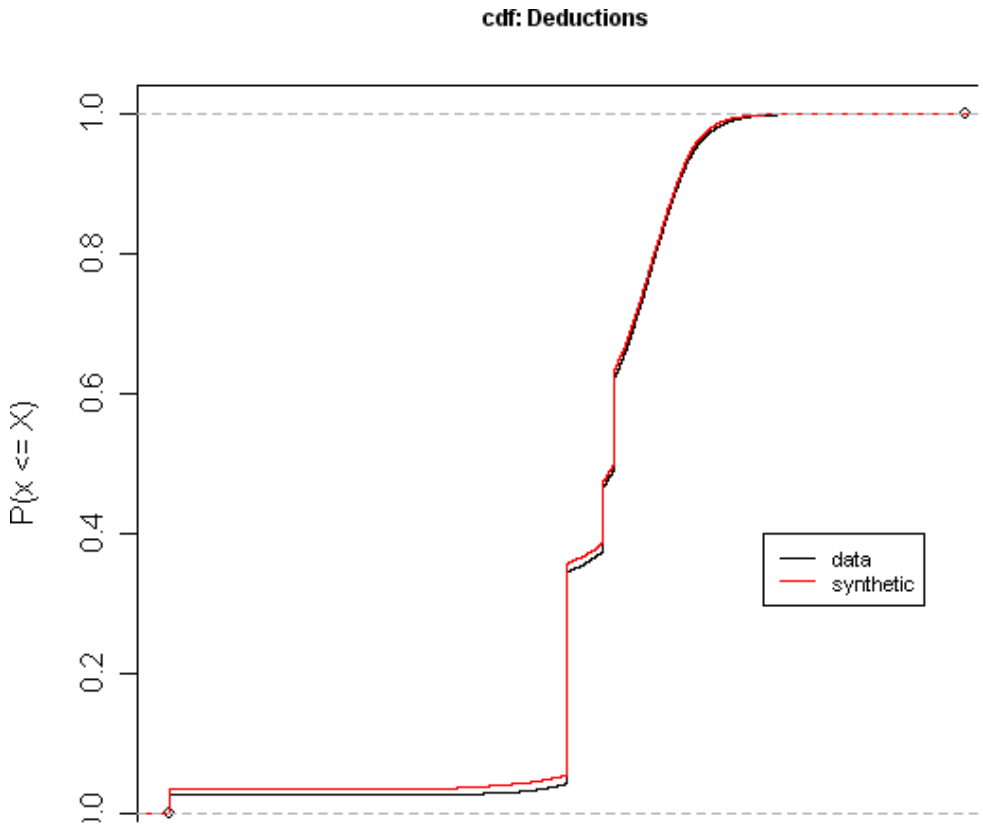

Figure 4.4 Empirical cumulative distribution functions of log(deductions) from original and synthetic Iowa income tax return data.

values on all variables, to create an entire synthetic set.

4.2 U.S. Census Bureau application

The U.S. Census Bureau collects and maintains data collected in surveys and censuses. To achieve its goal of simultaneously disseminating information while protecting confidentiality, the Census Bureau takes several approaches. Published statistics and summaries, tables, and subsamples with a limited number of variables and geographic information are among them. Users who wish to compute other statistics and perform their own analyses can apply for access to microdata through a Research Data Center or at the site of the Census Bureau itself. The process requires a proposal of research, oaths and contracts to protect confidentiality, and restriction to the physical location where research can be performed if proposals are accepted and access is granted. We suggest that the SDL methods described in this dissertation can be implemented on a number of data sets to produce releasable data to users, lessening the burden on users and on the Census Bureau itself.

Results from an application of generating synthetic values using quantile regression for veterans data in the American Community Survey at the U.S. Census Bureau are presented. In Section 4.2.1 we provide a description of the American Community Survey. The procedure used, including details on the quantile regression models used to generate synthetic data, is presented in Section 4.2.2. Initial results and some concerns are presented in Section 4.2.3.

4.2.1 American Community Survey

The U.S. Census Bureau administers a decennial census to provide population counts consistent with a constitutional mandate to apportion seats in the House of Representatives. The long form, which is sent to a random $\frac{1}{6}$ subsample of the households, has historically accompanied the decennial census to collect data on the social, economic, housing, and demographic characteristics of the population. With a growing population and increased needs for current and more frequent information about these characteristics, the American Community Survey (ACS) was designed. It is administered yearly and will replace the long form starting in 2010, thereby enabling the Census Bureau to provide pertinent and timely data products every year about communities with larger populations and every three years and about communities with smaller populations every 5 years. More detailed information is available in the ACS Handbook and a document describing the design and implementation

of the ACS (U.S. Census Bureau 2007).

4.2.2 Models and procedure

We apply the methods presented in Section 2.1 of this dissertation to ACS data on veterans. Specifically, we simulate synthetic values for age and wages using conditional quantile regression models. The values of age and wages in the data have distinct distributions for female and male respondents, so we consider separate models for the two groups. Based on discussions with members of the Statistical Research Division at the Census Bureau, some variables are included in the models to maintain important conditional distributions. Others are included based on empirical plots and correlations that indicate they help to characterize the distributions of age and wages.

We use a conditional model containing variables that reflect education levels (educ), current employment in the military (mil), social security income (ss), and fertility (fer) for female respondents. Define $x = \{\text{educ, mil, ss, fer}\}$ for male respondents and $x = \{\text{educ, mil, ss}\}$ for female respondents. The quantile regression model is

$$Q_{age}(\tau|x) = \xi_{age}(x, \beta_{age,\tau}) + \epsilon_{age,\tau}, \tag{4.2}$$

where ξ_{age} is a linear function of x and $\beta_{age,\tau}$, and $\epsilon_{age,\tau}$ represents independently and identically distributed random errors.

Values of wages are simulated using a conditional model containing age, commute time (commute), race group (race), retirement income (retire), social security income (ss), and two variables reflecting the amount of time spent at work (work_1 and work_2). Two considerations are made for wages. The first is due to a large number of records with recorded wages of zero. Rather than include all records in the estimation procedure, we first perform logistic regression to predict whether wages> 0 or wages$= 0$ in corresponding synthetic records. For records with predicted wages of zero, we assign a zero value to the corresponding synthetic record. Then we estimate parameters in a quantile regression model for wages using only data on records with the synthetic value of wages predicted to be positive.

The second consideration is due to the very wide range of values for wages. As in the Iowa Department of Revenue application, we apply a modified log-transformation to lessen

the effect of the highest values in the estimation:

$$l.wages = \begin{cases} log(wages), & wages \neq 0 \\ 0, & wages = 0. \end{cases} \tag{4.3}$$

With $x = \{$ age, commute, race, retire, ss, work$_1$, work$_2\}$ for both female and male respondents, the quantile regression model for log wages, or l.wages, is

$$Q_{l.wages}(\tau|x) = \xi_{l.wages}(x, \beta_{l.wages,\tau}) + \epsilon_{l.wages,\tau}, \tag{4.4}$$

where $\xi_{l.wages}$ is a linear function of x and $\beta_{l.wages,\tau}$, and $\epsilon_{l.wages,\tau}$ are independently and identically distributed random errors.

For both age and wages, we simulate values using the method described in Section 2.1.3. Values for $\tau_{age,j}$ and $\tau_{l.wages,j}$ are randomly selected for each record j from a Uniform$(0, 1)$ distribution. Due to computational contraints, rather than fit a quantile regression model for each randomly selected quantile, we first fit the models in Equations 4.3 and 4.4 for quantiles in the set $\tau = \{0.001, 0.01, 0.02, ..., 0.98, 0.99, 0.999\}$. Predictions for the randomly selected quantiles are interpolated using predictions for these quantiles.

For record j, we desire the predicted age at randomly selected quantile $\tau_{age,j}$. If predicted values $\hat{y}_{\tau_{age,a}}$ and $\hat{y}_{\tau_{age,b}}$ are computed at quantiles $\tau_{age,a}$ and $\tau_{age,b}$ from the set τ directly above and below the randomly selected quantile, then we can interpolate between the predicted values to obtain the desired synthetic value as follows:

$$\hat{y}_{\tau_{age,j}} = \tau_{age,j}\hat{y}_{\tau_{age,a}} + (1 - \tau_{age,j})\hat{y}_{\tau_{age,b}}. \tag{4.5}$$

A similar method is used to obtain synthetic values for l.wages. Synthetic values for wages can be obtained by taking the exponential of the computed synthetic wages. Results are presented in Section 4.2.3.

4.2.3 Results

Recall, the procedure was implemented separately on records for female and male respondents. We compare marginal distributions of age and wages in the original and synthetic data

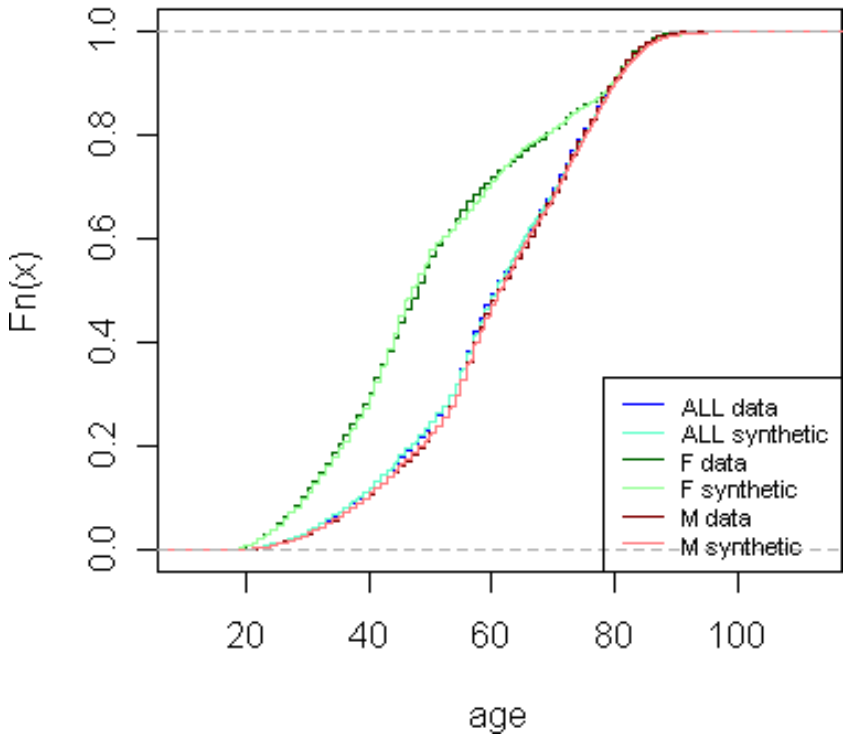

Figure 4.5 Empirical cumulative distribution functions of age in original and synthetic American Community Survey (ACS) veterans data.

using plots of their empirical cumulative distributions in Figures 4.5 and 4.6. Marginally, the distributions of age and l.wages in the data are fairly well preserved in the distribution of the synthetic values. This occurs when all records are lumped together and also within female and male groups.

In an effort to assess the extent to which this method preserved the conditional distribution of l.wages given age, commute time (commute), and two variables reflecting the amount of time spent at work, work_1 and work_2, linear regression estimates, standard errors, and R^2 values from the model $l.wages = f(\text{age}, \text{work}_1, \text{work}_2, \text{commute}, \gamma) + \epsilon$ are presented in Table 4.1. Parameter estimates for the linear regression model are similar. However, standard errors of the estimates computed using the synthetic data set are higher than in the original data set. This makes sense, since the method used to generate synthetic values introduces extra randomness to both age and l.wages values which affects the accuracy, or standard

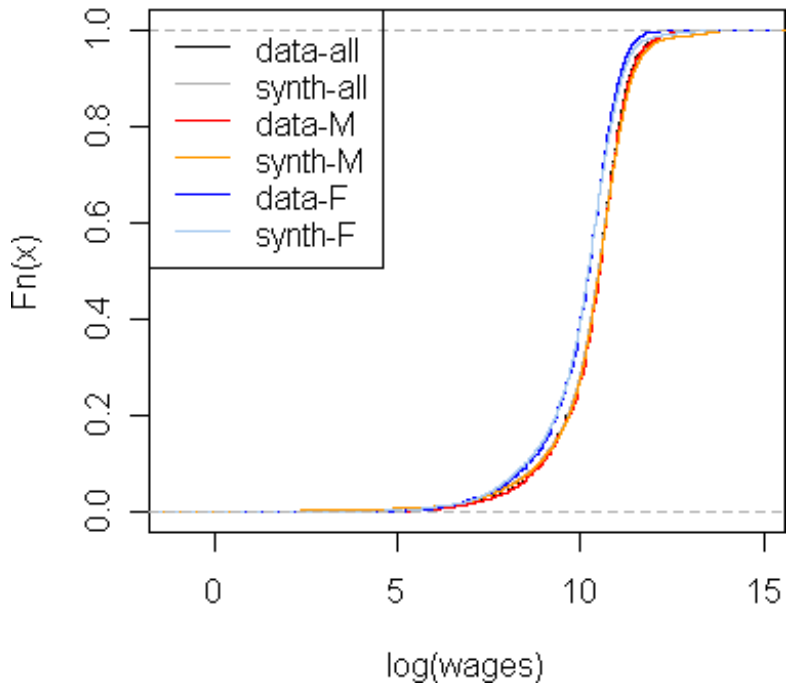

Figure 4.6 Empirical cumulative distribution functions of log(wages) from original and synthetic American Community Survey (ACS) veterans data.

Table 4.1 Linear regression results for predicting l.wages from synthetic and original American Comunity Survey veterans data sets.

	coefficient	estimate	$s.e.(\hat{\gamma})$
Original data	$\hat{\gamma}_{age}$	0.0027	0.00036
	$\hat{\gamma}_{work_1}$	0.0304	0.00027
	$\hat{\gamma}_{work_2}$	0.0432	0.00027
	$\hat{\gamma}_{commute}$	0.0033	0.00014
	R^2	0.46	
Synthetic data	$\hat{\gamma}_{age}$	-0.0007	0.00033
	$\hat{\gamma}_{work_1}$	0.0289	0.00037
	$\hat{\gamma}_{work_2}$	0.0411	0.00036
	$\hat{\gamma}_{commute}$	0.0033	0.00018
	R^2	0.30	

errors, of estimates based on those values. The coefficient of age actually is far (in terms of standard errors) from the original estimated coefficient. The value of R^2 is also a lot smaller. One could ask what happened to these coefficients. One possibility is that the statistical disclosure limitation method was not applied to all the variables used in the model used to assess data utility. As a result, the simulated ages and wages (or l.wages) do not necessarily match well with other variables that were not generated conditional on their values. As a consequence, a recommendation can be made to fully create the synthetic data set for all variables to be released or to only release variables created in a consistent manner.

A practical concern that the Census Bureau has about releasing synthetic data maintaining the consistency of values in individual records. Consider age and Veteran Period of Service (VPS), for instance. An example of one such inconsistency is a record containing a synthetic value of 17 for age, say, and a value corresponding to World War II for VPS. We examine the distributions of values for age within VPS categories. Synthetic age values rarely fall outside the range of data values, thus few inconsistencies exist. To ensure the resulting synthetic age values fall within the range of values in the data, age values were generated using the quantile regression method within VPS categories. The box plots in Figure 4.7 show the ranges of age values within VPS categories in the original and synthetic data. If synthetic age values that fall outside of the range of values in the original data are truly nonsensical, then further work needs to be done to correct for this. If values outside the range in the data are plausible, then it might be acceptable to allow them in the synthetic

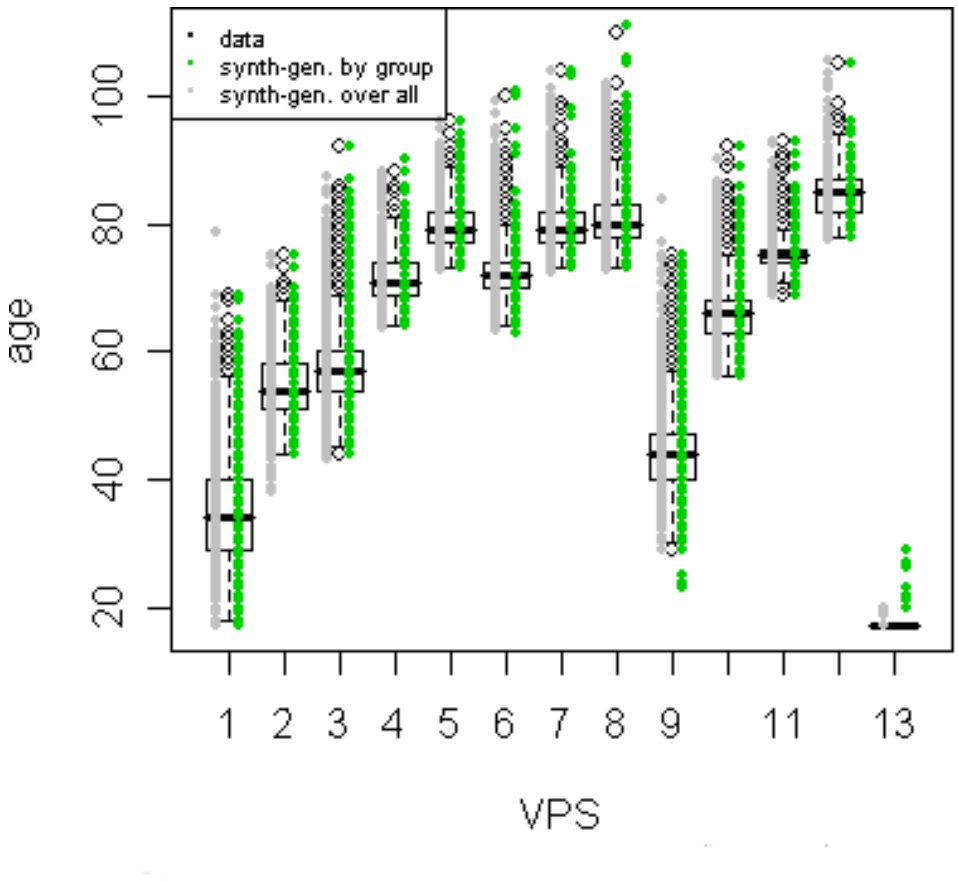

Figure 4.7 Box plots of age within Veteran Period of Service (VPS) from original and synthetic American Community Survey (ACS) veterans data.

data set. Further investigation to this topic should be considered.

The results presented in this section highlight the appeal of quantitative measures of data utility as well as disclosure risk. How similar or different should point estimates be between the synthetic and orginal data sets? How much added variability is enough? How much is too much? These are questions that can be answered by comparing risk and utility of the data set that results from applying different SDL methods. While we do not have quantitative measures to compare the synthetic and original data sets, we can see that such measures would be very helpful in assessing any SDL method used to protect confidential data.

In Section 4.3, the proposed method is applied to a public use microdata sample available through the Census Bureau's website. In this application, we study the entire proposed SDL method in order to implement our developments to measure disclosure risk in the resulting synthetic data for different types of intruders and targets.

4.3 Public Use Microdata Sample application

The U.S. Census Bureau makes microdata sets available through what are called Public Use Microdata Samples (PUMS). PUMS include microdata files from the American Community Survey (ACS). Identifying information is removed, geographic locations are reported in broad regions. Because of the limitations due to confidentiality, we could not continue to work on the ACS data set described in Section 4.2 outside of the Census Bureau facility. We substitute a similar data set taken from the PUMS data for an application in measuring disclosure risk.

Access to PUMS can be gained through the Census Bureau website (www.census.gov) and through an application called DataFerret, available for download through the Census website (dataferrett.census.gov) as well. Ferret is an acronym for Federated Electronic Research, Review, Extraction and Tabulation Tool. The application is free and easy to use. It provides access to samples of data from surveys that are collected by the Census Bureau. Included is the ACS among several others. To use the DataFerret application, a user must submit information including name and email address. Once access is granted (almost immediately) the user may select which survey to download data from and which variables from the survey s/he wants information on. Geographical information is limited to broad regions (northeast, midwest, south, west, Puerto Rico), divisions (Puerto Rico, New England, Middle Atlantic, etc.), and states. Other variables include information on housing and population characteristics as well as replicate weights for variables.

We use a PUMS data set as a test bed to implement the SDL procedure proposed in Chapter 2, compute risk as formulated in Chapter 3 for three types of intruder and three targets, and to evaluate the data utility in the resulting synthetic data set. The PUMS data has already had SDL methods applied to it to protect confidentiality before being released to the public. Hence, in our application, we assume some hypothetical situations in which records would be sensitive if more detailed information was known. We also ignore any po-

Table 4.2 Variables in the U.S. Census Bureau Public Use Microdata Sample (PUMS) sample.

SCHL	educational attainment
VPS	veteran period of service
AGE	age
RET	retirement income
SS	social security income
SSI	supplemental social security income
TAX	property tax paid
WAGE	wages earned.

tential SDL methods that have been applied to the data set prior to our using it and assume values on records are the orignial data.

The PUMS sample used here includes over 30,000 records on respondents in Iowa with the following variables: age, retirement income, social security income, supplemental security income, property tax paid (categorized), wages earned, educational attainment, citizenship, sex, and veteran period of status. The are presented in Table 4.3.1. We suppose the scenario is that a researcher wants access to the agency's original data set to study income relative to age, educational attainment, and veteran period of service. Other relationships are important to maintain as well.

4.3.1 Models and procedure

In this section the proposed procedure is implemented to create synthetic data. Measures of disclosure risk presented in Chapter 3 are applied.

Educational attainment and veteran period of service are considered nonsensitive and available to the intruder. Their values are released essentially unperturbed. Educational attainment (SCHL) has 17 levels. The value 0 is recorded for respondents who are less than 3 years old, 1 is recorded for respondents with no education, values increase from 1 to 16 with grade levels and degrees. Level 16 corresponds to a doctorate degree. Veteran period of service (VPS) has 13 levels with values 0 through 12. A value of 0 corresponds to a respondent who is not a veteran. The majority of respondents have recorded value 0. Levels from 1 to 12, correspond to veteran respondents who are veterans of the Gulf War

Table 4.3 Recode variables Veteran's Period of Service (VPS) and educational attainment (SCHL) from U.S. Census Bureau Public Use Microdata Sample (PUMS) data set into new variable SCHL/VPS.

$SCHL/VPS$	VPS	$SCHL$
1	0	0
2	0	1
3	0	2
4	0	3
5	0	4
6	0	5
7	0	6
8	0	7
9	0	8
10	0	9
11	0	10
12	0	11
13	0	≥ 12
14	≥ 1	≤ 8
15	≥ 1	$9, 10, 11$
16	≥ 1	≥ 12

(1), the Gulf War and Vietnam (2), Vietnam only (3), etcetera, to before World War II only (12). Detailed variable definitions are available through the PUMS data dictionary at http://www.census.gov/acs/www/Downloads/PUMSDataDict06.pdf.

When VPS greater than 0, unique records exist with respect to SCHL and VPS levels. Thus, prior to release, these variables are recategorized. There are 16 resulting categories listed in Table 4.3. The resulting categories contain a sufficient number of records to be released safely. Again, since this is a public use microdata sample, there is not enough identifying information for the records to actually be sensitive. However, we can imagine that if these records were from a specific geographic location, and if that information was released with the data set, then the unique records would be sensitive.

Synthetic data for release is generated for AGE, RET, and SS using quantile regression predictions at randomly drawn quantiles. Hot deck imputation and rank swapping are used to generate values for SSI, TAX, and WAGE. The illustration in Figure 4.8 represents the

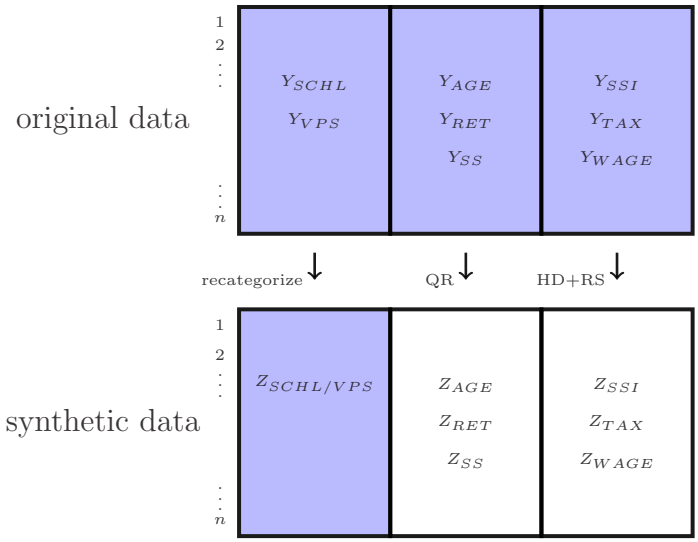

Figure 4.8 An illustration of the original U.S. Census Bureau Public Use Microdata Sample (PUMS) data set and STEP ONE of the creation of the synthetic data set. The original data set contains variables Y. In step one, the nonsensitive variables Y_{SCHL} and Y_{VPS} are recategorized to $Z_{SCHL/VPS}$ before being copied into the synthetic data set.

first stage of the procedure, in which values on SCHL and VPS are recategorized and copied into the synthetic data set. Details about the quantile regression models and predicted values follow in Section 4.3.1.1. Hot deck imputation and rank swapping details are presented in Section 4.3.1.2.

4.3.1.1 Quantile regression models for PUMS application

Synthetic values for AGE, RET, and SS are generated using conditional quantile regression models. Models and predicted values are estimated in a sequential fashion in order to retain the joint conditional distribution of AGE, RET, and SS, given SCHL and VPS. First, models for AGE are estimated from the original data, given SCHL and VPS. Predicted values $\widehat{AGE}_{\tau_{AGE}}$ are computed using values of SCHL and VPS on each record and the regression coefficients estimated at quantiles τ_{AGE}. Second, models for RET are estimated using the original data, given AGE, SCHL, and VPS at quantiles τ_{RET}. Predicted values $\widehat{RET}_{\tau_{RET}}$ are computed using synthetic values $\widehat{AGE}_{\tau_{AGE}}$ and original values of SCHL and VPS on each record. Third, models for SS are estimated from the original data, given RET, AGE, SCHL, and VPS. Predicted values $\widehat{SS}_{\tau_{SS}}$ are computed using regression estimates at

quantiles τ_{SS}, synthetic values $\widehat{AGE}_{\tau_{AGE}}$ and $\widehat{RET}_{\tau_{RET}}$, and original values SCHL and VPS. The conditional quantile regression models are written as:

$$\begin{aligned} Y_{AGE,\tau_{AGE}} &= (Y_{SCHL}\ Y_{VPS})\beta_{AGE,\tau_{AGE}} + \epsilon_{AGE,\tau_{AGE}} \\ Y_{RET,\tau_{RET}} &= (Y_{SCHL}\ Y_{VPS}\ Y_{AGE})\beta_{RET,\tau_{RET}} + \epsilon_{RET,\tau_{RET}} \\ Y_{SS,\tau_{SS}} &= (Y_{SCHL}\ Y_{VPS}\ Y_{AGE}\ Y_{RET})\beta_{SS,\tau_{SS}} + \epsilon_{SS,\tau_{SS}}. \end{aligned} \tag{4.6}$$

with errors ϵ_k independently and identically distributed with a normal distribution at quantile τ with zero mean and variance $\omega(\tau_k)^2 = \frac{\tau_k(1-\tau_k)}{f^2(F^{-1}(\tau_k))}$, for variable $k = AGE, RET, SS$.

As discussed in Section 4.2.2, the quantile regression models in Equation 4.6 of this application are fit at each quantile in the set $\tau = \{0.001, 0.01, ..., 0.99, 0.999\}$. Predicted values at randomly drawn quantiles from the interval $(0, 1)$ are interpolated using predicted values at quantiles from this set. Specifically, a random quantile $\tau^*_{AGE,j}$ is drawn for each record $j = 1, ..., n$. Values $\tau_{a,j}$ and $\tau_{b,j}$ from the set τ are directly above and below $\tau^*_{AGE,j}$. The synthetic predicted value, $\hat{y}_{AGE,\tau^*_{AGE}}$, is obtained by interpolating between predicted values at $\tau_{a,j}$ and $\tau_{b,j}$, $\hat{y}_{AGE,\tau_a}$ and $\hat{y}_{AGE,\tau_b}$. This can be written as follows:

$$\begin{aligned} \hat{y}_{AGE,\tau_{a,j}} &= (y_{SCHL,j}\ y_{VPS,j})\hat{\beta}_{AGE,\tau_a} \text{ and} \\ \hat{y}_{AGE,\tau_{b,j}} &= (y_{SCHL,j}\ y_{VPS,j})\hat{\beta}_{AGE,\tau_b}. \end{aligned} \tag{4.7}$$

Using these two predicted values, the predicted value $\hat{y}_{AGE,\tau^*_{AGE,j}}$ is computed using the following interpolation:

$$\hat{y}_{AGE,\tau^*_{AGE,j}} = \tau^*_{AGE,j}\hat{y}_{AGE,\tau_{a,j}} + (1 - \tau^*_{AGE,j})\hat{y}_{AGE,\tau_{b,j}}. \tag{4.8}$$

This is done to generate a synthetic value of AGE for each record in the synthetic set.

Interpolation is also used to generate synthetic values for RET conditional on SCHL, VPS, and AGE, and similarly, to generate synthetic values for SS conditional on SCHL, VPS, AGE, and RET. Expressions for this are shown here:

$$\begin{aligned} \hat{y}_{RET,\tau^*_{RET,j}} &= \tau^*_{RET,j}\hat{y}_{RET,\tau_{a,j}} + (1 - \tau^*_{RET,j})\hat{y}_{RET,\tau_{b,j}} \text{ and} \\ \hat{y}_{SS,\tau^*_{SS,j}} &= \tau^*_{SS,j}\hat{y}_{SS,\tau_{a,j}} + (1 - \tau^*_{SS,j})\hat{y}_{SS,\tau_{b,j}}, \end{aligned} \tag{4.9}$$

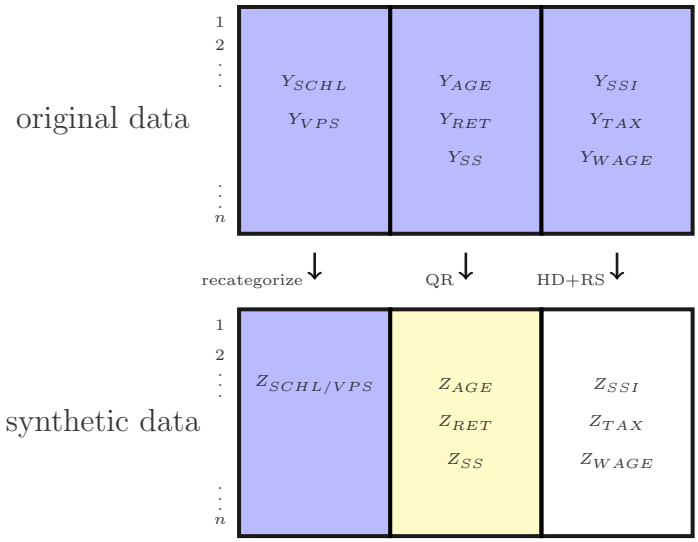

Figure 4.9 An illustration of the original U.S. Census Bureau Public Use Microdata Sample (PUMS) data set and STEP TWO of the creation of the synthetic data set. In step two, the sensitive variables Y_{AGE}, Y_{RET}, and Y_{SS} are replaced in the synthetic data set using quantile regression predictions for variables Z_{AGE}, Z_{RET}, and Z_{SS}.

where

$$\hat{y}_{RET,\tau^*_{RET,j}} = (y_{SCHL,j}\ y_{VPS,j}\ \hat{y}_{AGE,\tau^*_{AGE,j}})\hat{\beta}_{RET,\tau^*_{RET,j}} \text{ and}$$

$$\hat{y}_{SS,\tau^*_{SS,j}} = (y_{SCHL,j}\ y_{VPS,j}\ \hat{y}_{AGE,\tau^*_{AGE,j}}\ \hat{y}_{RET,\tau^*_{RET,j}})\hat{\beta}_{SS,\tau^*_{SS,j}}.$$

The illustration in Figure 4.9 reflects the SDL procedure and values in the synthetic data set up to this point with original values shaded in blue and synthetic quantile regression predicitons shaded in yellow.

4.3.1.2 Hot deck and rank swapping procedure for PUMS application

Hot deck imputation and rank swapping are used to generate values for supplemental security income (SSI), property tax (TAX), and wage income (WAGE) for the synthetic data set. In the hot deck portion of the procedure, matching between each synthetic record j and all original records $i = 1, ..., n$ is performed using the Mahalanobis distance based on AGE and RET values. Original record i with the smallest distance $d(i, j)$ to synthetic record j is selected as the matching record. In practice, to add more variablility to the synthetic data set, we randomly choose from among records with distance in the lowest ten percent of

matching records.

Rather than impute all values for SSI, TAX, and WAGE into synthetic record j, further perturbation is introduced by performing rank swapping on the matching original record i. We compute the sample rank of values $Y_{SSI,i}, Y_{TAX,i}$, and $Y_{WAGE,i}$ from the matching recrod i, and call them $r_{SSI,i}, r_{TAX,i}$, and $r_{WAGE,i}$. These ranks are swapped with random ranks $r^*_{SSI,i}, r^*_{TAX,i}$, and $r^*_{WAGE,i}$. Values on records with the random ranks are imputed into synthetic record j. Ranks $r^*_{SSI,i}, r^*_{TAX,i}$, and $r^*_{WAGE,i}$ are randomly selected from uniform distributions over intervals centered at the sample ranks of values in matching record i, i.e.

$$\begin{aligned} r^*_{SSI,i} &\sim \text{Uniform}(r_{SSI,i} - \delta, r_{SSI,i} + \delta_1), \\ r^*_{TAX,i} &\sim \text{Uniform}(r_{TAX,i} - \delta_2, r_{TAX,i} + \delta_2), \text{ and} \\ r^*_{WAGE,i} &\sim \text{Uniform}(r_{WAGE,i} - \delta_3, r_{WAGE,i} + \delta_3). \end{aligned}$$

The values of δ_1, δ_2, and δ_3 in this application were set to 20. As discussed in Section 2.3, increasing the value of δ would increase distortion from the original data causing an undesireable decrease in data utility but a simultaneous decrease disclosure risk. It would be an interesting study to determine how much of a change in δ would correspond to a practical change in both data utility and disclosure risk. This is out of the scope of the work done in this application.

At this point, the synthetic data set is complete. The illustration in Figure 4.10 shows that the variables Y_{SCHL} and Y_{VPS} have been recategorized into $Z_{SCHL/VPS}$, synthetic values for Y_{AGE}, Y_{RET}, and Y_{SS} have been generated using quantile regression predictions Z_{AGE}, Z_{RET}, and Z_{SS} (shaded in yellow), and hot deck imputation and rank swapping have been used to produce values Z_{SSI}, Z_{TAX}, and Z_{WAGE} as synthetic versions (shaded in red) of Y_{SSI}, Y_{TAX}, and Y_{WAGE}.

4.3.2 Results

Results from implementing the proposed procedure are presented in this section. First, empirical plots and regression analyses are presented to assess data utility. Then disclosure risk is assessed using the framework and formulation for risk measures presented in

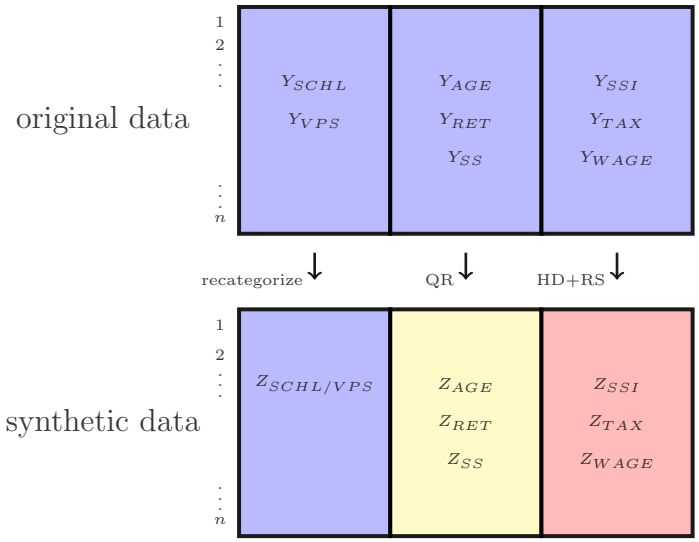

Figure 4.10 An illustration of the original U.S. Census Bureau Public Use Microdata Sample (PUMS) data set and STEP THREE of the creation of the synthetic data set. In step three, the sensitive variables Y_{SSI}, Y_{TAX}, and Y_{WAGE}. are replaced in the synthetic data set using hot deck imputation and rank swapping for variables Z_{SSI}, Z_{TAX}, and Z_{WAGE}.

Chapter 3. Risk is assessed for three types of intruders and three targets. The results are summarized in the next sections.

4.3.2.1 Data utility

In this section, empirical plots of densities are presented to assess the degree to which marginal distributions are preserved in the synthetic data set. Regression analyses are useful to assess conditional distributions.

Marginal distributions

Densities of synthetic and original values of AGE, RET, SS, SSI, TAX, and WAGE are presented in Figure 4.11. Synthetic values on AGE, RET, and SS were generated in step two of the procedure using the quantile regression models in Equations 4.10 and 4.11. Values of SSI, TAX, and WAGE were generated in step three of the procedure using hot deck imputation and rank swapping. Plots of empirical densities for each of these six variables allow us to assess data utility with respect to the marginal distributions of each variable in

the original and synthetic data sets.

The plots in Figure 4.11 show that the distribution of synthetic values of AGE, RET, SS, and SSI look similar to distributions in the original data. There are discrepancies in the frequency of -1 and 0 values in SS and SSI. There are obvious discrepancies between the distribution of TAX and WAGE between the original and synthetic data sets. To better visualize the differences in these variables and also to allow a closer look at the distributions of RET, SS, and SSI, the distributions of log values are plotted. Values on RET, SS, SSI, and WAGE have a lower bound of 0, with -1 values indicating the respondent's age is less than 15. This log transformation assigns the following values to $log(variable)$:

$$\begin{array}{rll} & = -1 & \text{when } variable = -1 \\ log(variable) & = 0 & \text{when } variable = 0 \\ & = log(variable) & \text{when } variable > 0. \end{array}$$

Densities of the log values can be seen in Figures 4.12 through 4.15. Since the magnitude of AGE and TAX values are relatively small and neither distribution has a huge density of values around -1 or 0, we do not transform these values.

In Figures 4.12 through 4.15, the marginal distributions of variables are presented in the density plots of log values. They show that RET, SS, and SSI distributions in the original and synthetic data sets are similar. Though, in the right tail of SS, plots show a lower frequency of records with values of log(SS) between 8 and 10, corresponding to SS values between 3,000 and 22,000. A notable difference between original and synthetic distributions is found in the density plot of WAGE values in Figure 4.15. Prior to the log transformation, the density curve implies that there is a large discrepancy between the distribution of WAGE values in the original and synthetic data sets, roughly at WAGE values of 0, 3,000, and 13,000. When the density of the log values is plotted, it seems that the discrepancies of frequencies are not as extreme as they appear in the original plot. However, when the 0 and -1 values are eliminated, the synthetic data show extra peaks in the upper end of the distribution, with peaks at log(WAGE) vlues around 8 and 9. These values correspond to WAGE values of almost 3,000 and about 13,000. Also, the right tail of the WAGE distribution in the synthetic data set does not extend as far as in the original data set. The implications of this will depend on the purpose intended by the data user.

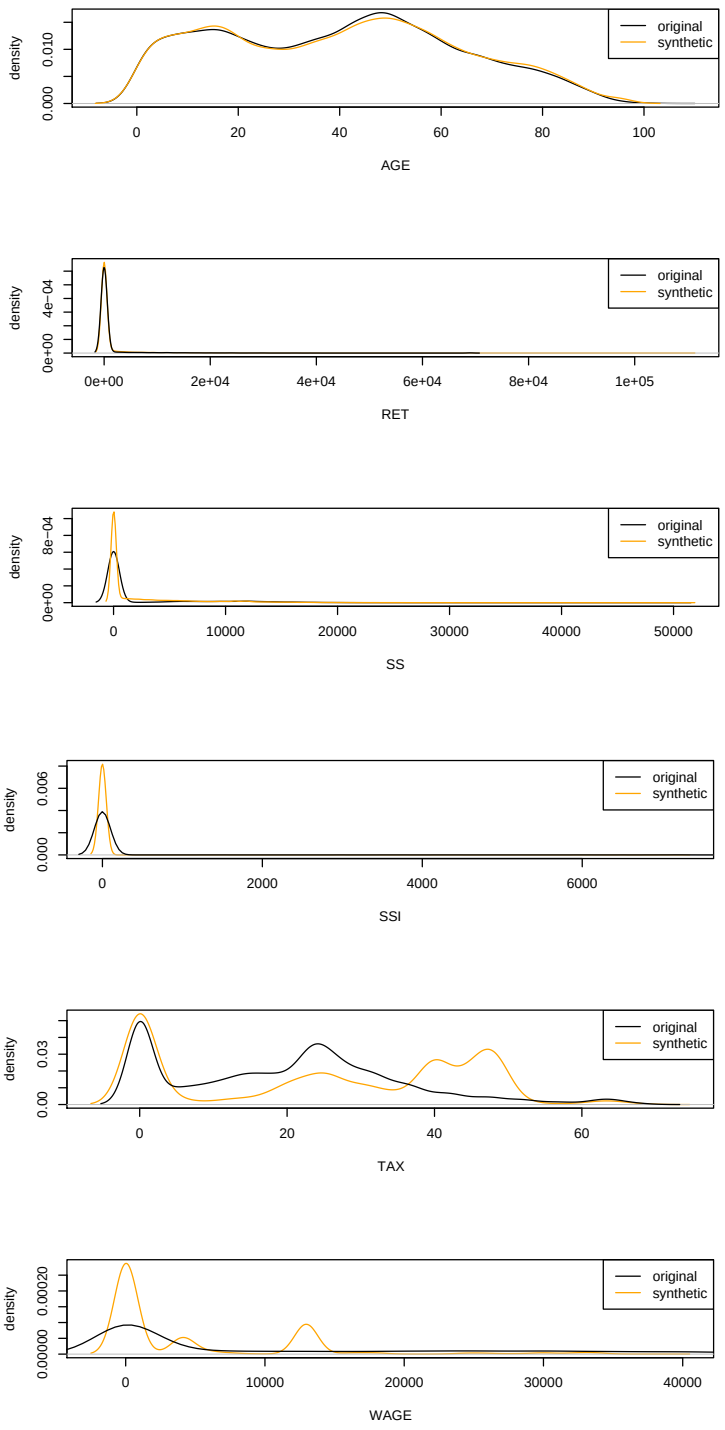

Figure 4.11 Empirical densities for AGE, RET, SS, SSI, TAX, and WAGE from original and synthetic U.S. Census Bureau Public Use Microdata Sample (PUMS) data.

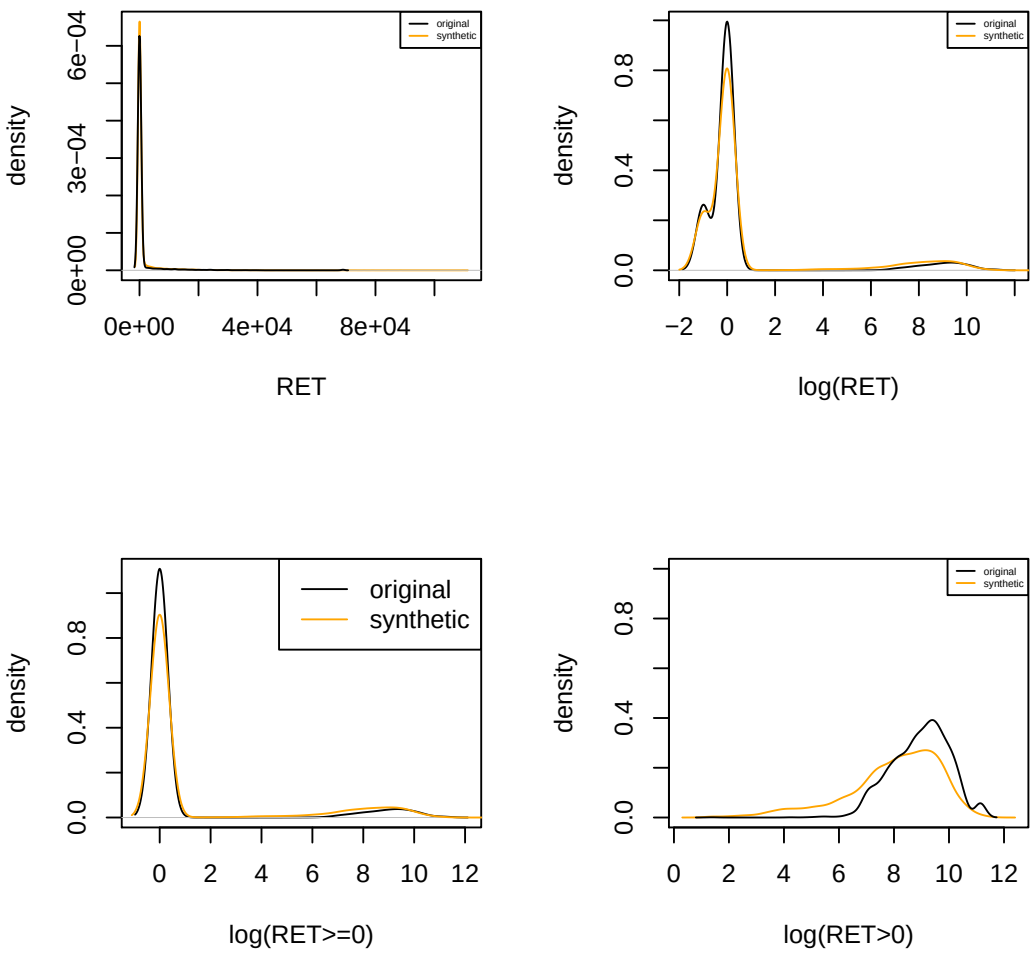

Figure 4.12 Empirical densities for RET on the original and log scale from original and synthetic U.S. Census Bureau Public Use Microdata Sample (PUMS) data.

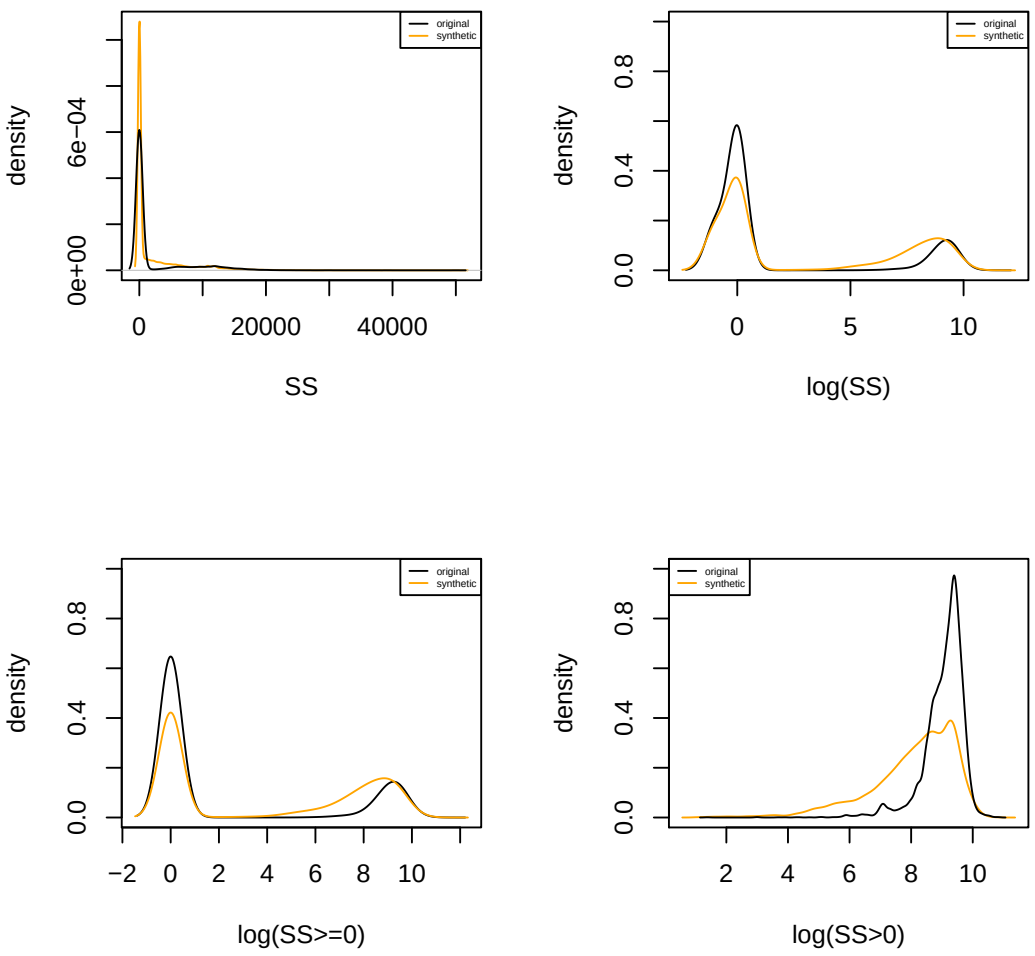

Figure 4.13 Empirical densities for SS on the original and log scale from original and synthetic U.S. Census Bureau Public Use Microdata Sample (PUMS) data.

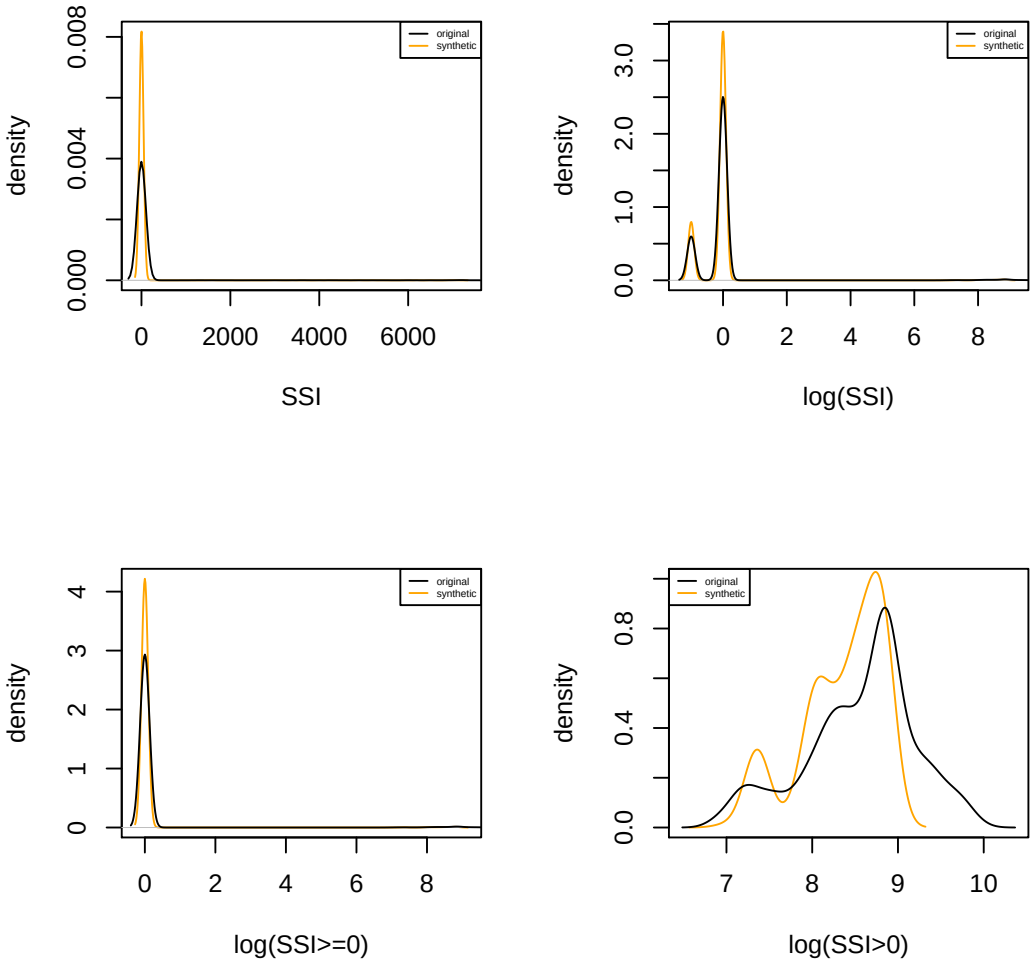

Figure 4.14 Empirical densities for SSI on the original and log scale from original and synthetic U.S. Census Bureau Public Use Microdata Sample (PUMS) data.

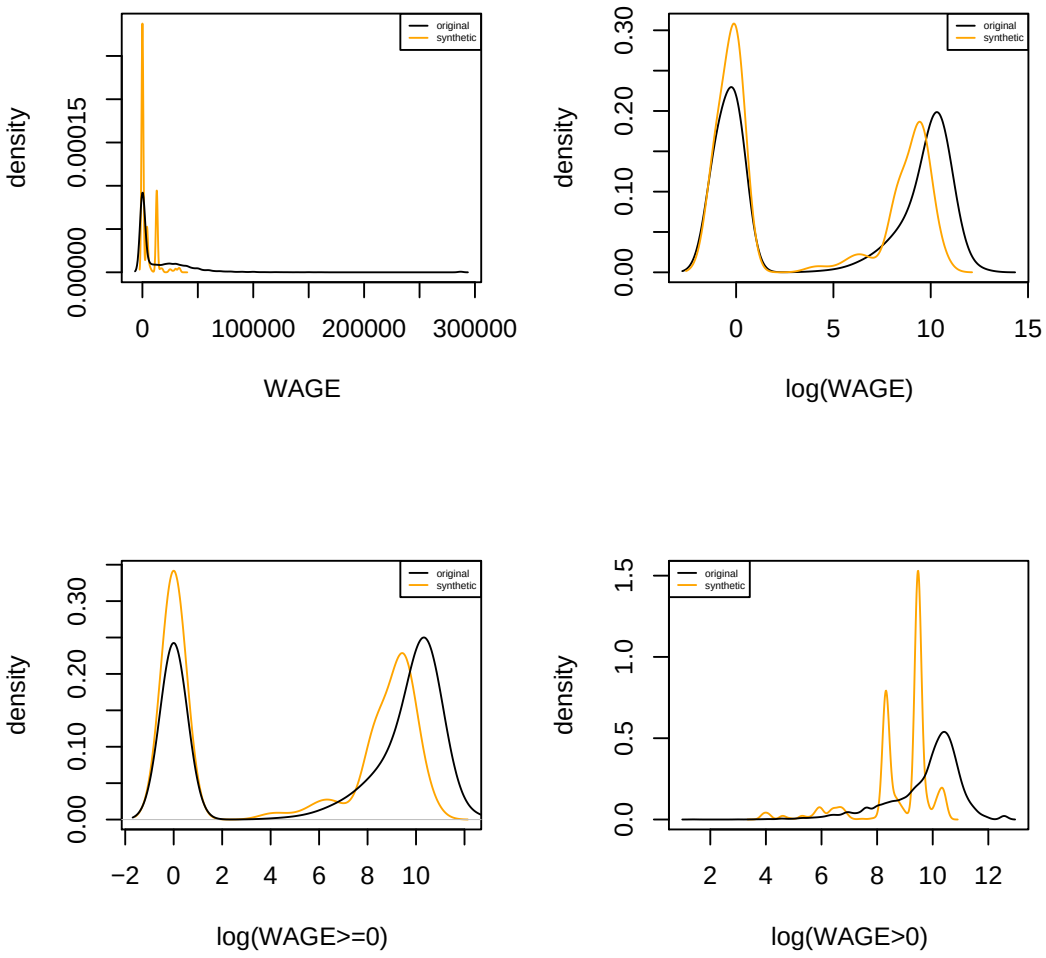

Figure 4.15 Empirical densities for WAGE on the original and log scale from original and synthetic U.S. Census Bureau Public Use Microdata Sample (PUMS) data.

Conditional distributions

To assess the conditional distributions in the synthetic data set compared to those in the original data set, we compare linear regression results. In particular, we examine the distribution of responses conditional on SCHL/VPS category to assess how well the conditional relationships in the original data set are represented in the synthetic data set. Other analyses and models can be considered. If an agency had an idea of specific analyses of interest to a user, results from those particular analysis could be compared between the original and synthetic data sets to assess data utility.

The regression estimates and standard errors from the linear regression model of AGE on SCHL/VPS category are presented in Table 4.4. These results show that the synthetic values of AGE provide essentially the same estimates and standard errors when regressed on SCHL/VPS category as the original data do. Notice that the standard errors of each estimated coefficient are slightly lower for the synthetic data than for the original data. This could be due to synthetic AGE values being quantile regression estimates conditional on SCHL and VPS values. At particular values of SCHL and VPS, similar quantile regression predictions of AGE will have been made, decreasing the variation among values in each SCHL/VPS category. Notice also, that the R^2 value for the synthetic data is slightly lower, so although the variation of AGE values within SCHL/VPS categories is slightly lower, the added variability (from predictions being made at random quantiles) in the overall data may not be accounted for by this linear regression model.

The results in Table 4.4 show that the synthetic data for AGE values retain the conditional relationship with SCHL/VPS categories that appears in the original data. In Tables 4.5 through 4.8, we present results from regression models of RET, SSI, TAX, SS, and WAGE on SCHL/VPS category. Figures 4.16 through 4.21 provide plots to visualize the estimates and standard errors using "standardized intervals." The standardized intervals are centered at each parameter estimate $\hat{\beta}$ divided by its standard error $s.e.(\hat{\beta})$, with form $(\hat{\beta}/s.e.(\hat{\beta}) - 2, \hat{\beta}/s.e.(\hat{\beta}) + 2)$. In each plot, the points corespond to the value $(\hat{\beta}/s.e.(\hat{\beta})$ estimated using the original, or real data ($\mathbf{r}$) and the synthetic data ($\mathbf{s}$). Intervals surrounding $\mathbf{r}$ and $\mathbf{s}$ of the same color belong to estimates of the same parameter.

The tables and plots show that relationships in the data are maintained to a varying degree. By examining the tables we see that, oddly, in some cases, the R^2 increases in the synthetic set. This may not be as odd as an initial reaction though. The response variables

Table 4.4 Linear regression results: $Y_{AGE} = Y_{SCHL/VPS}\beta_{AGE} + \epsilon_{AGE}$.

	ORIGINAL		SYNTHETIC	
SCHL/VPS	estimate	s.e.	estimate	s.e.
(intercept)	1.00	0.540	0.98	0.533
2	7.65	0.767	7.45	0.756
3	7.45	0.651	7.07	0.642
4	15.14	0.784	14.62	0.774
5	33.76	0.703	35.56	0.694
6	28.04	0.841	28.89	0.829
7	33.59	0.814	35.34	0.803
8	32.93	0.807	34.38	0.796
9	48.50	1.092	47.27	1.077
10	50.24	0.577	48.12	0.569
11	44.32	0.697	43.82	0.688
12	41.50	0.625	42.42	0.617
13	43.75	0.585	43.96	0.577
14	72.70	1.085	61.62	1.070
15	60.19	0.674	61.59	0.665
16	57.58	0.859	56.68	0.847
	$R^2 = 0.4679$		$R^2 = 0.4622$	

we have regressed on SCHL/VPS category all have values generated using values of those categories. Thus, in the synthetic data set, as an artifact of the synthetic data method, SCHL/VPS categories directly account for more of the variation in the response variable values. By looking at the tables alone, it seems the situation is somewhat dire–that the synthetic data do not produce the same regression estimates as the original data. Most of the plots, however, show that the standardized intervals overlap. For instance, in Figure 4.16, we see that **r** and **s** pairs are close and that their intervals overlap. Most of the intervals do not contain 0 (vertical line at 0 on the x-axis), indicating parameter estimates are significant in both data sets. This gives us confidence that the synthetic data set, used in such an anlysis, will lead to similar conclusions as the original data.

We noted earlier that densities of TAX and WAGE in the synthetic data do not align nicely with those from the original data. Discrepancies can be seen in the conditional relationship of both of these variables with SCHL/VPS categories too. For the most part, in Figures 4.19 and 4.21, standardized intervals from the synthetic and original data sets do not overlap. For TAX, the implications may be severe since most estimates from the original data are not signifcantly different from zero but corresponding estimates from the synthetic data set are. This warrants further investigation, especially if the agency is aware that a user of this synthetic data set is interested in studying property tax values with respect to these SCHL/VPS categories.

In further analyses, we compare results from regression analyses between AGE and the other variables. The results are presented in Table 4.9. Results for each of five regression analyses are contained in this table and in Figures 4.22 through 4.25.

4.3.2.2 Disclosure risk assessment for synthetic PUMS data set

Using the developments presented in Chapter 3, we assess disclosure risk in the synthetic PUMS data set. Disclosure risk is considered for an SDL intruder with detailed knowledge of the SDL procedure used, an average intruder with prior knowledge including regression estimates from the original data set, and a naive intruder with knowledge only from the released synthetic data set. Each intruder knows the original values on particular target records. These include a target record that is unique in the data set, a target record that is rare, and a target record that is common with respect to the SCHL/VPS category.

Values on the target records are provided in Table 4.10. The left column in the table

Table 4.5 Linear regression results: $logY_{RET} = Y_{SCHL/VPS}\beta_{RET} + \epsilon_{RET}$.

	ORIGINAL		SYNTHETIC	
SCHL/VPS	estimate	s.e.	estimate	s.e.
(intercept)	-1.00	0.074	-1.00	0.086
2	0.13	0.105	0.19	0.122
3	0.03	0.089	0.01	0.103
4	0.15	0.108	0.24	0.124
5	0.95	0.097	1.20	0.112
6	1.28	0.115	1.83	0.133
7	1.46	0.112	2.16	0.129
8	1.42	0.111	2.16	0.128
9	1.93	0.150	2.41	0.173
10	1.82	0.079	2.30	0.091
11	1.75	0.096	2.24	0.111
12	1.53	0.086	2.18	0.099
13	1.70	0.080	2.25	0.093
14	3.66	0.149	4.18	1.172
15	3.41	0.093	3.21	0.107
16	3.83	0.118	3.01	0.136
	$R^2 = 0.1219$		$R^2 = 0.1013$	

Table 4.6 Linear regression results: $logY_{SSI} = Y_{SCHL/VPS}\beta_{SSI} + \epsilon_{SSI}$.

	ORIGINAL		SYNTHETIC	
SCHL/VPS	estimate	s.e.	estimate	s.e.
(intercept)	-1.00	0.032	-1.00	0.020
2	0.18	0.045	0.09	0.028
3	0.02	0.038	0.01	0.023
4	0.18	0.046	0.12	0.029
5	0.66	0.041	0.60	0.026
6	1.25	0.049	0.97	0.031
7	1.28	0.048	1.11	0.030
8	1.28	0.047	2.98	0.030
9	1.27	0.064	1.00	0.040
10	1.20	0.034	1.00	0.021
11	1.14	0.041	1.00	0.026
12	1.08	0.037	1.00	0.022
13	1.04	0.034	1.00	0.021
14	1.24	0.064	1.00	0.040
15	1.09	0.039	1.00	0.025
16	1.05	0.050	1.00	0.032
	$R^2 = 0.1386$		$R^2 = 0.3860$	

Table 4.7 Linear regression results: $Y_{TAX} = Y_{SCHL/VPS}\beta_{TAX} + \epsilon_{TAX}$.

	ORIGINAL		SYNTHETIC	
SCHL/VPS	estimate	s.e.	estimate	s.e.
(intercept)	18.22	0469	17.55	0.257
2	1.01	0.666	1.80	0.365
3	3.19	0.565	10.76	0.310
4	3.42	0.681	-17.55	0.373
5	0.01	0.611	11.75	0.335
6	-0.17	0.730	-15.44	0.400
7	-0.73	0.707	23.77	0.388
8	-1.88	0.701	8.10	0.384
9	-2.03	0.949	-4.58	0.520
10	0.15	0.501	26.54	0.275
11	-0.41	0.605	7.68	0.332
12	-0.59	0.543	27.61	0.298
13	6.58	0.508	-17.55	0.278
14	-1.53	0.942	8.72	0.517
15	1.64	0.585	-3.36	0.321
16	7.39	0.746	3.43	0.409
	$R^2 = 0.0334$		$R^2 = 0.8159$	

Table 4.8 Linear regression results: $Y_{TAX} = Y_{SCHL/VPS}\beta_{TAX} + \epsilon_{TAX}$.

	ORIGINAL		SYNTHETIC	
SCHL/VPS	estimate	s.e.	estimate	s.e.
(intercept)	18.22	0469	17.55	0.257
2	1.01	0.666	1.80	0.365
3	3.19	0.565	10.76	0.310
4	3.42	0.681	-17.55	0.373
5	0.01	0.611	11.75	0.335
6	-0.17	0.730	-15.44	0.400
7	-0.73	0.707	23.77	0.388
8	-1.88	0.701	8.10	0.384
9	-2.03	0.949	-4.58	0.520
10	0.15	0.501	26.54	0.275
11	-0.41	0.605	7.68	0.332
12	-0.59	0.543	27.61	0.298
13	6.58	0.508	-17.55	0.278
14	-1.53	0.942	8.72	0.517
15	1.64	0.585	-3.36	0.321
16	7.39	0.746	3.43	0.409
	$R^2 = 0.0334$		$R^2 = 0.8159$	

Table 4.9 Linear regression results: $Y = Y_{AGE}\beta + \epsilon$.

	ORIGINAL			SYNTHETIC		
SCHL/VPS	estimate	s.e.	R^2	estimate	s.e.	R^2
(intercept)	-898.041	54.701	0.0513	47.720	49.156	0.0174
RET	47.838	1.171		24.899	1.064	
(intercept)	4.820	9.536	0.0033	41.147	4.688	0.0426
SSI	2.074	0.204		-0.206	0.102	
(intercept)	18.987	0.171	0.0014	22.468	0.216	0.0061
TAX	0.025	0.004		0.065	0.005	
(intercept)	-2585.007	40.517	0.36	-35.505	46.729	0.0981
SS	114.320	0.867		58.658	1.012	
(intercept)	10,887.794	307.445	0.0126	1109.428	78.385	0.0809
WAGE	130.545	6.581		88.522	1.697	

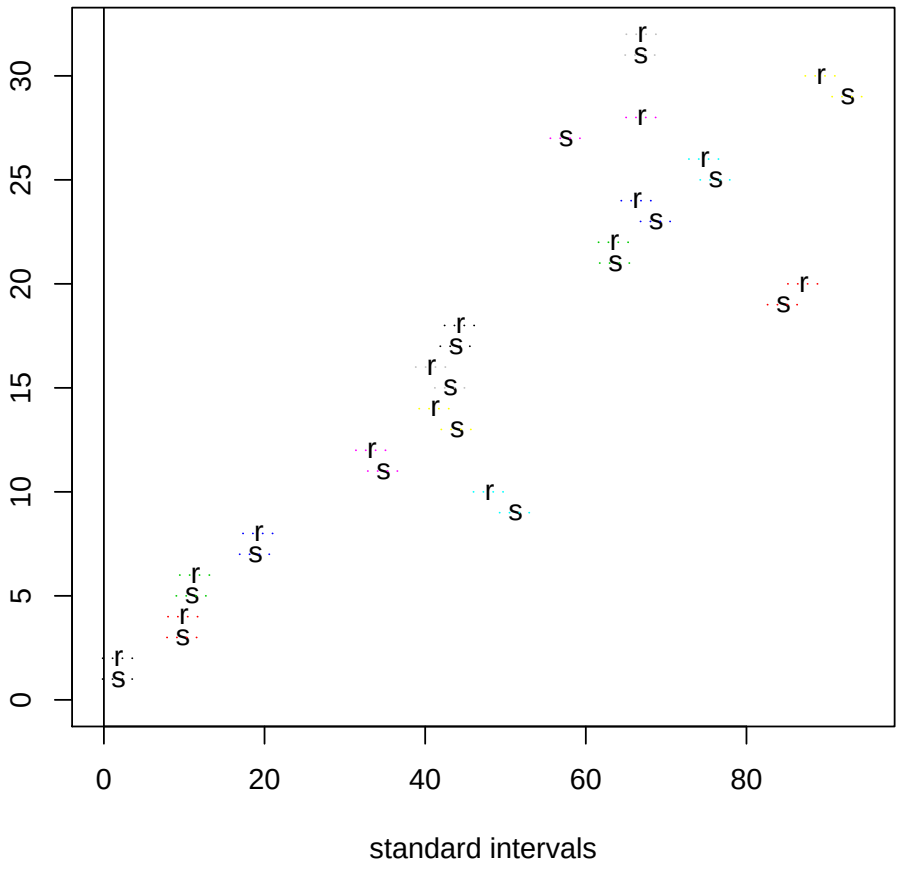

Figure 4.16 $Y_{AGE} = Y_{SCHL/VPS}\beta_{AGE} + \epsilon_{AGE}$: standardized intervals for $\hat{\beta}_{AGE}$.

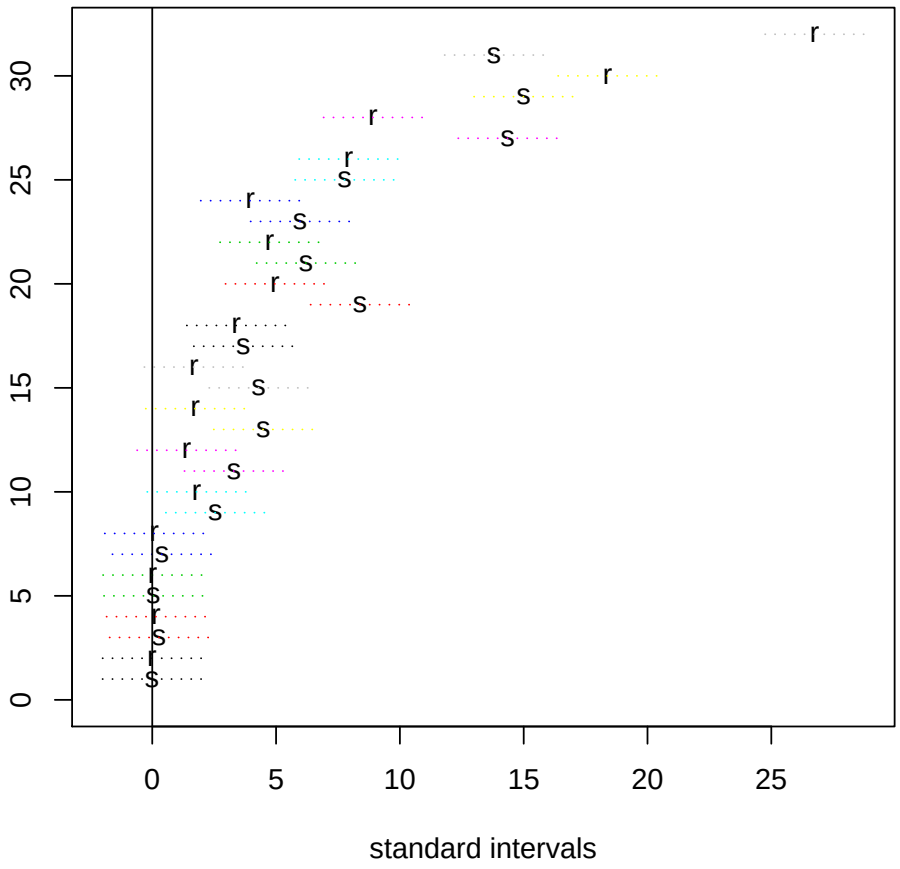

Figure 4.17 $Y_{RET} = Y_{SCHL/VPS}\beta_{RET} + \epsilon_{RET}$: standardized intervals for $\hat{\beta}_{RET}$.

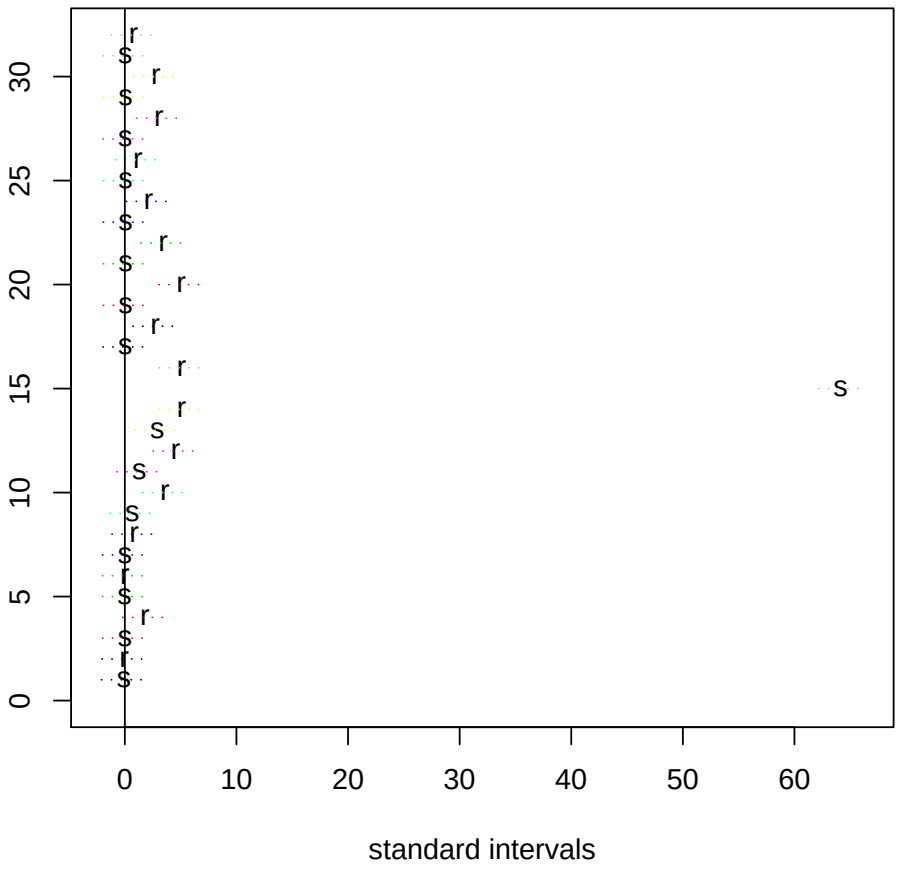

Figure 4.18 $Y_{SSI} = Y_{SCHL/VPS}\beta_{SSI} + \epsilon_{SSI}$: standardized intervals for $\hat{\beta}_{SSI}$.

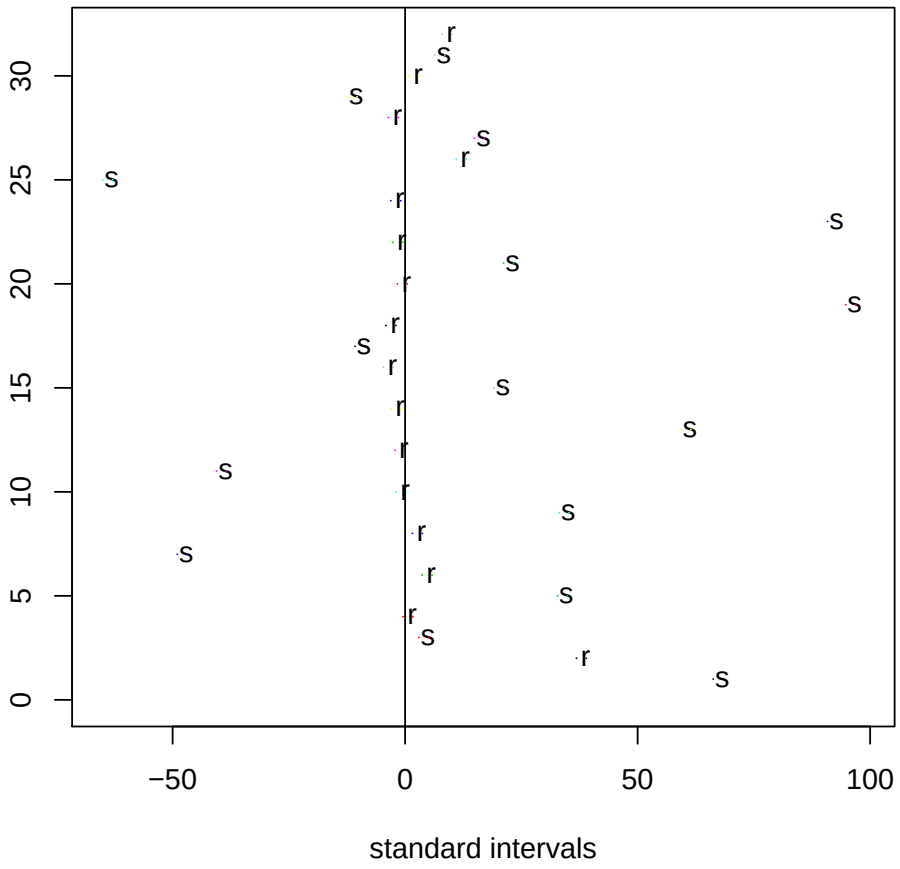

Figure 4.19 $Y_{TAX} = Y_{SCHL/VPS}\beta_{TAX} + \epsilon_{TAX}$: standardized intervals for $\hat{\beta}_{TAX}$.

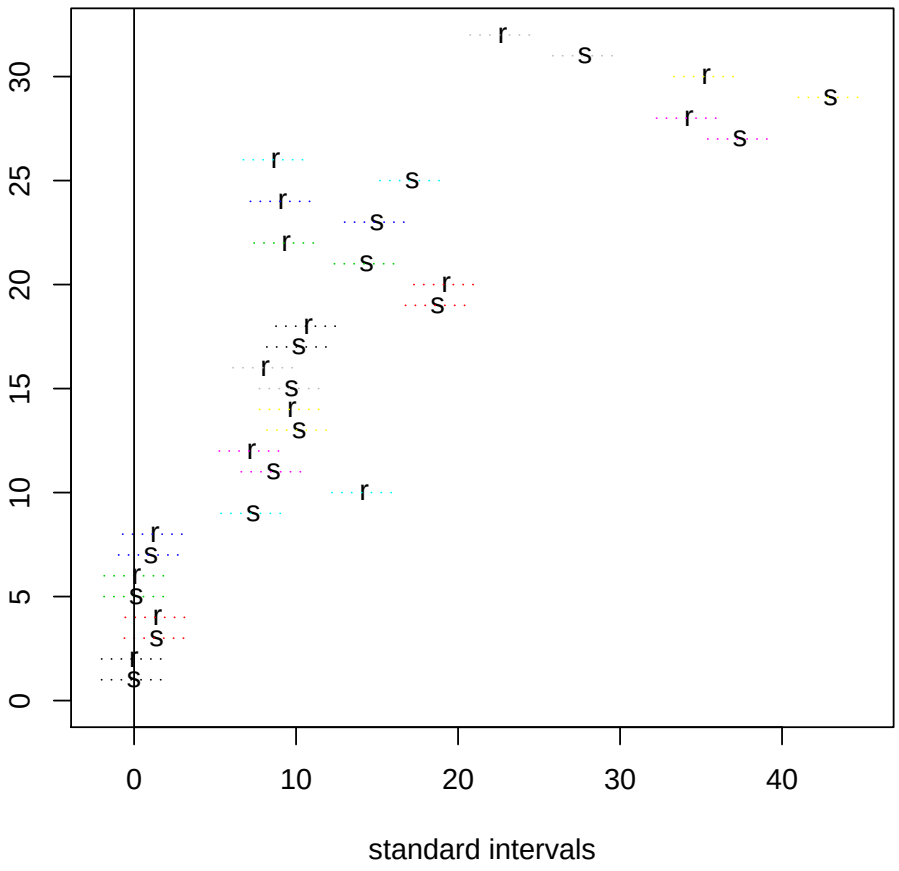

Figure 4.20 $Y_{SS} = Y_{SCHL/VPS}\beta_{SS} + \epsilon_{SS}$: standardized intervals for $\hat{\beta}_{SS}$.

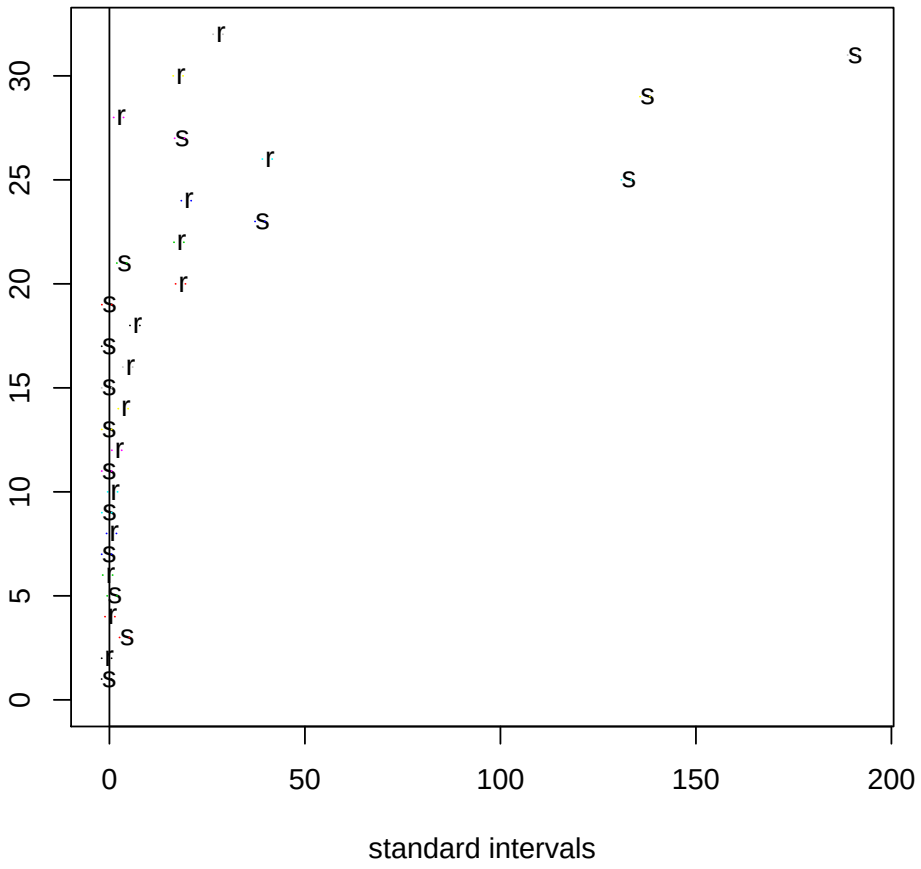

Figure 4.21 $Y_{WAGE} = Y_{SCHL/VPS}\beta_{WAGE} + \epsilon_{WAGE}$: standardized intervals for $\hat{\beta}_{WAGE}$.

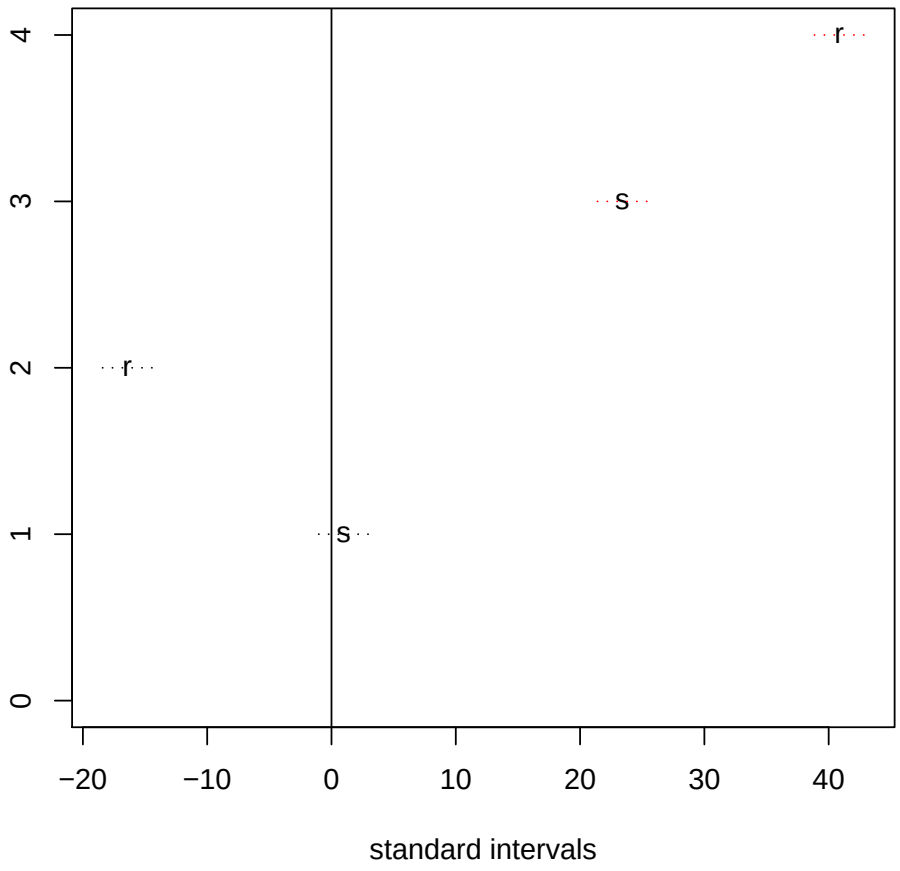

Figure 4.22 $Y_{RET} = Y_{AGE}\beta_{RET} + \epsilon_{RET}$: standardized intervals for $\hat{\beta}_{RET}$.

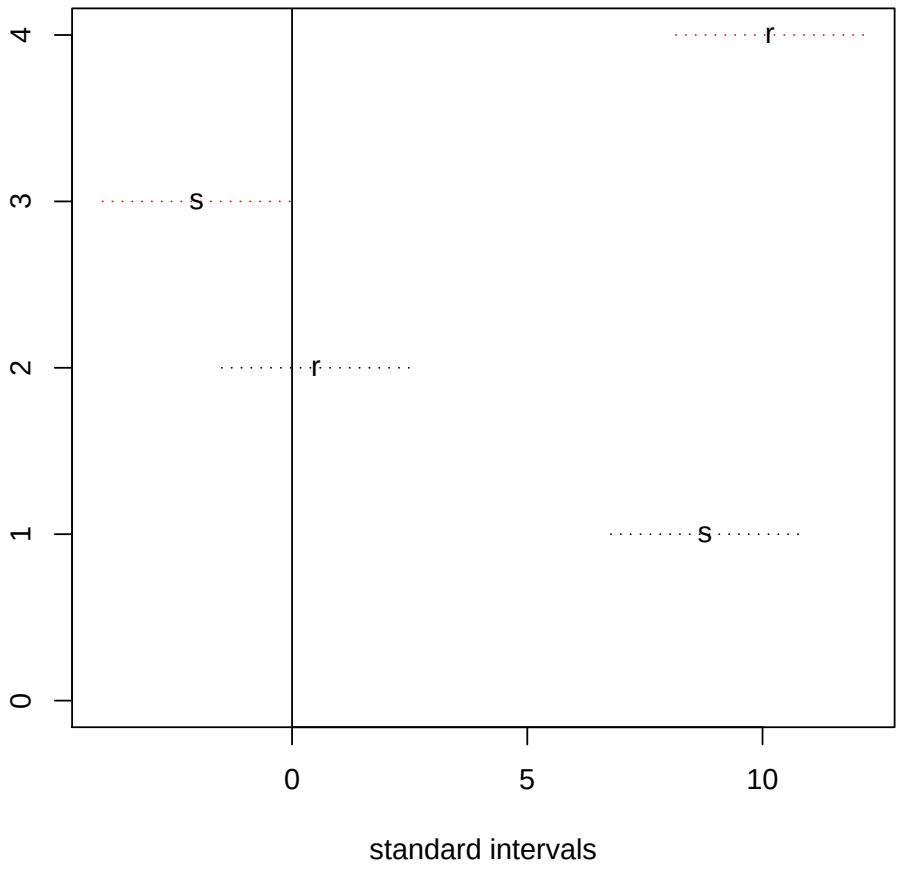

Figure 4.23 $Y_{SSI} = Y_{AGE}\beta_{SSI} + \epsilon_{SSI}$: standardized intervals for $\hat{\beta}_{SSI}$.

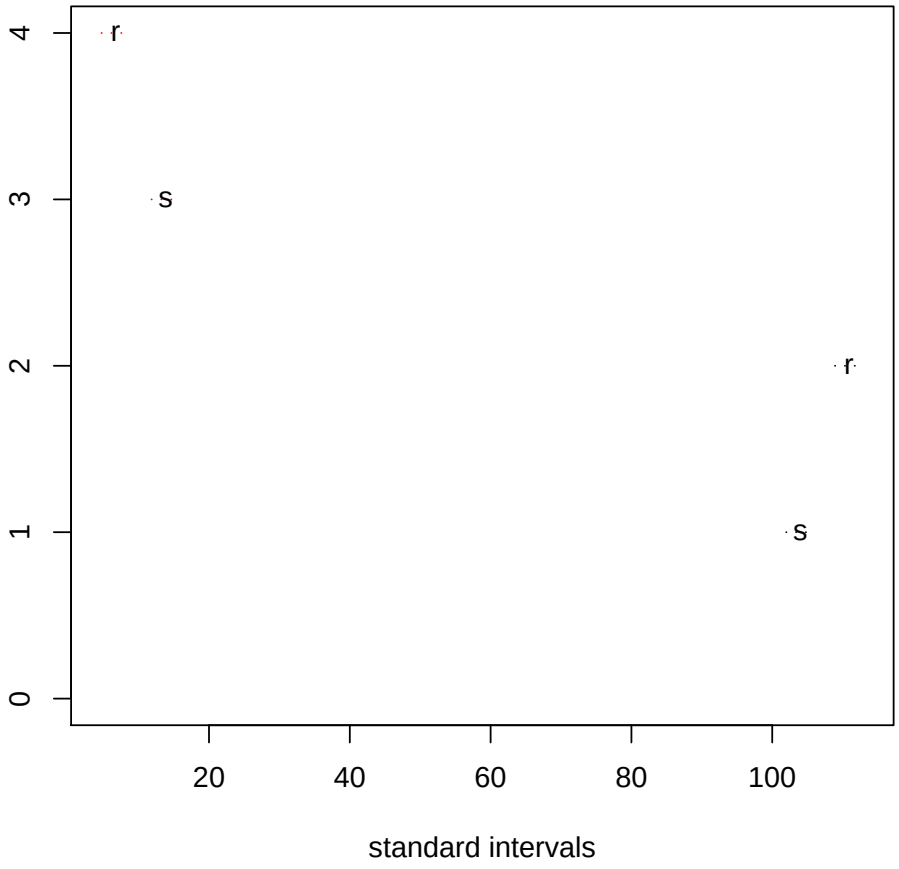

Figure 4.24 $Y_{TAX} = Y_{AGE}\beta_{TAX} + \epsilon_{TAX}$: standardized intervals for $\hat{\beta}_{TAX}$.

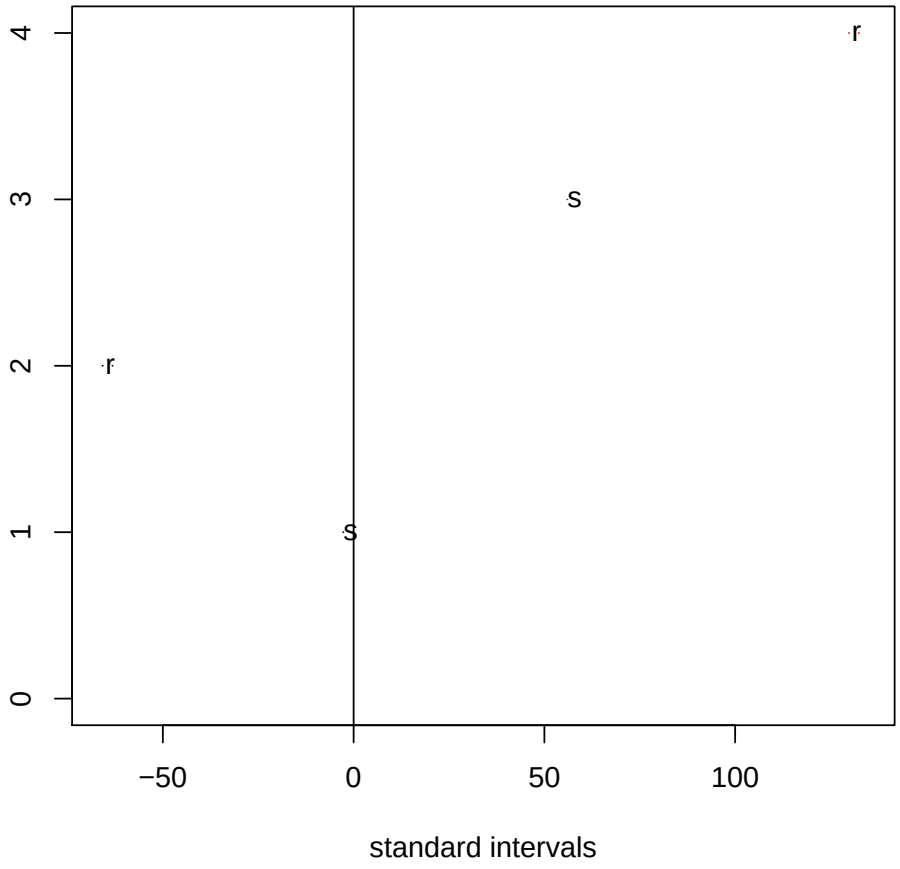

Figure 4.25 $Y_{SS} = Y_{AGE}\beta_{SS} + \epsilon_{SS}$: standardized intervals for $\hat{\beta}_{SS}$.

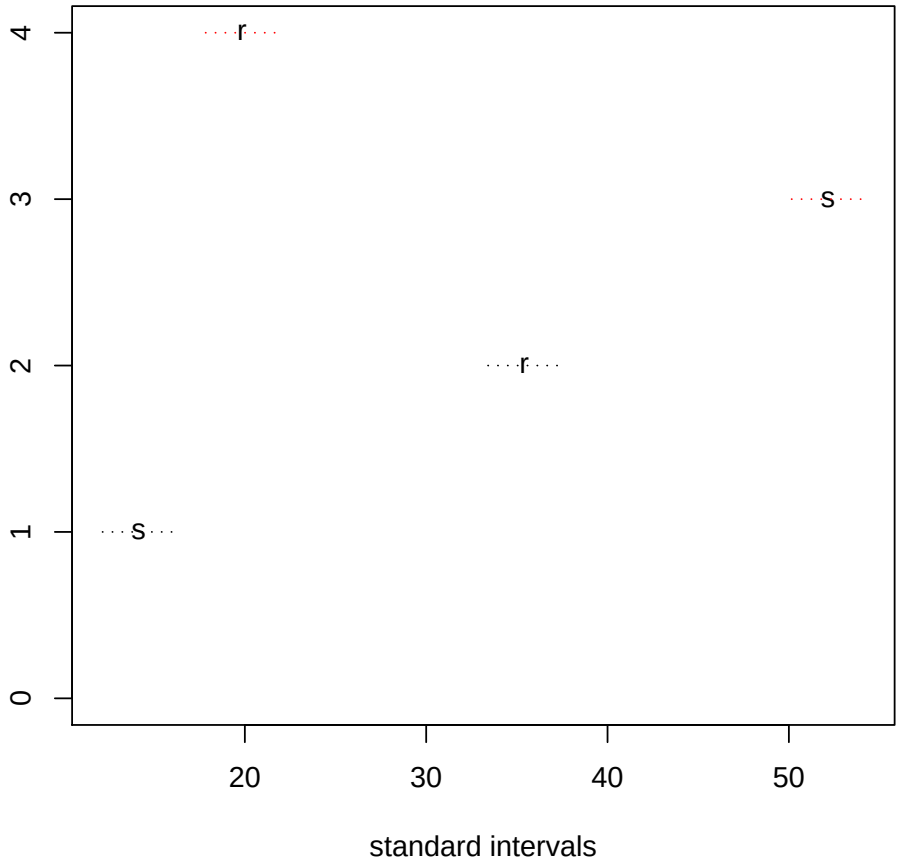

Figure 4.26 $Y_{WAGE} = Y_{AGE}\beta_{WAGE} + \epsilon_{WAGE}$: standardized intervals for $\hat{\beta}_{WAGE}$.

Table 4.10 Target records in the original and synthetic U.S. Census Burea Public Use Microdata Sample (PUMS) data sets. Values for the unique, rare, and common targets are give. Original values are on the left. Synthetic values are on the right.

		UNIQUE		RARE		COMMON	
SDL set	variable	original	synthetic	original	synthetic	original	synthetic
Ad	SCHL	16	-	3	-	9	-
	VPS	11	-	8	-	1	-
	SCHL/VPS	16	16	14	14	15	15
Ap	$z_1 = AGE$	73	80	90	42	32	50
	$z_2 = RET$	69,000	0	0	0	0	0
	$z_3 = SSI$	0	0	0	0	0	0
	$z_4 = TAX$	47	0	14	24	28	0
Up	$z_5 = SS$	20,200	0	15,700	10,799	0	5952
	$z_6 = WAGE$	0	28,000	0	4,100	287,000	34,500

includes values from original records for each target. The unique target is the only record in the data set with recorded values corresponding to the respondent having attained a doctorate degree (SCHL=16) and veteran period of service occuring during the peace-time between the Korean War and World War II (VPS=11). The unique target contains high values of RET at \$69,000 and SS at \$20,200 in the original data set. The rare record is one of three with educational attainment of 5^{th} or 6^{th} grade (SCHL=3) and veteran period of service during World War II (VPS=8). From these three records, the one chosen as the target contains recorded AGE and SS values higher that the other two rare records. The common record is one of 112 records containing educational attainment of high school graduation (SCHL=9) and veteran period of service during the Gulf War (VPS=1). The record chosen as the target has an exceptionally high value of WAGE at \$287,000 compared to the other 111 records in which all recorded WAGE values are less than \$100,000.

Values on the synthetic records that result from the SDL procedure are presented in the right columns in Table 4.10. None of the target's synthetic records correspond very closely with the original records. Variable SCHL/VPS was not perturbed, so it matches exactly. The other variables were perturbed. This indicates that disclosure risk is low for each of the targets. To quantify this, we compute disclosure risk according to the procedure outline in Chapter 3.

Table 4.11 Disclosure risk for the synthetic U.S. Census Burea Public Use Microdata Sample (PUMS) data set for three targets (unique, rare, and common) for three intruders (SDL, average, naive).

		$Pr(J = j\|t, Z)$		
		UNIQUE	RARE	COMMON
	SDL	0.00259	7.5×10^{-13}	0.00058
intruder	*average*	0	0	0
	naive	0	0	0
	naive*	0	0	0.11217

Recall from Equation 3.1 that the disclosure risk for each record j is written:

$$Pr(J = j|t, z) = \frac{A_j B_j C_j \frac{1}{n_t}}{\sum_{i=1}^{n} A_i B_i C_i \frac{1}{n_t}},$$

for ever $j = 1, ..., n$, where n_t is the number of records with $t^{Ad} = z_j^{Ad}$. The resulting disclosure risk for each target and intruder is presented in Table 4.11. Disclosure risk is set to 0 for any released record with SCHL/VPS category not equal to the target's value on this category. For records released with the same category as the target, disclosure risk is computed. The results are summarized for each intruder type below.

SDL intruder

To compute the disclosure risk attributed to the SDL intruder, we assume that specific knowledge of the SDL procedure is available to the intruder and that the intruder chooses to use this information to compute the probability of identification. The SDL intruder knows the original values on the targets for variables SCHL, VPS, AGE, RET, SSI, and TAX. S/he knows that SCHL and VPS are recategorized into broader categories defined in Table 4.3, and thus knows the released value of VPS/SCHL for each target. It is known that values for AGE, RET, and SS were generated using quantile regression predictions from the models in Equation 4.6. The SDL intruder also knows that values in the released set, $z_{SSI,j}^{Ap}$, $z_{TAX,j}^{Ap}$, and $z_{WAGE,j}^{Up}$, were generated using hot deck imputation and rank swapping.

Recall that SDL risk is computed component-wise according to the form of components C_{SDL}, B_{SDL}, and A_{SDL}, as presented in Equations 3.25, 3.28, 3.29 and 3.32. The resulting disclosure risk presented in Table 4.11 is promising. The disclosure risk from the SDL in-

truder for the unique target is only 0.259 percent. The disclosure risk for the common target and the rare target are even less. Given the information and knowledge available to the SDL intruder, it seems fortunate that the risk is not higher. It seems that the discrepancies between the synthetic variables and the targets values confuse the SDL intruder and decrease the disclosure risk.

To examine disclosure risk overall, we can examine the values of $Pr_{SDL}(J = j|t, Z)$ computed for every record. Records with a different value of SCHL/VPS category as the target are set to zero, since SCHL/VPS categoryy are the variables in Ad, or those available to the intruder that do not get perturbed. For records with the same value of SCHL/VPS category recorded as each of the unique, rare, and common targets, the disclosure risk or probabilitiy of identification $Pr_{SDL}(J = j|t, Z)$ is computed. In Figure 4.27, three plots represent risk associated with each of the three targets. The top plot shows the disclosure risk, or probability of identification computed for records in SCHL/VPS category with the unique target. To link the unique target, it is assumed that among records with SCHL/VPS equal to the target's value, the probability of identification for each record is computed conditional on the record belonging to the target. High values of probability would indicate a high probability of a record belonging to the target.

To determine the likelihood of linking a record to the target, the value of its identification probability is considered relative to the other record in the category. For example, if the identification probability associated with a particular record is computed to be 0.25, this may be too high in an absolute sense, but if all of the records in the categrory are computed to have the same probability of belonging to the target, then the intruder may be no more inclined to choose this record over any other. The purpose of the plot in Figure 4.27 is to show that the identification probability, or risk, for the unique target is not only low on an absolute scale at 0.00259 or 0.259%, but that its risk is also low relative to the other records an intruder is attempting to link with the unique target.

It may be of interest to the agency to examine records with high disclosure risk, conditioning on the record belonging to the target, even when the record is not the target's record. In the uniqe's SCHL/VPS category, the record corresponding to the largest probability of identification has a value of 0.00911 or 0.91%. Although this record is not the synthetic version of the target's record, the agency may want to examine the record to determine if a false identification with the target would be likely and the effects that a false identification would have. The agency may decide to further alter such a record in order to decrease its

disclosure risk to the level of the other records.

The center plot of Figure 4.27 displays the disclosure risk associated with the 345 records in the same SCHL/VPS category as the rare target. Recall that the rare target shares original values of SCHL and VPS with two other records. The probability of identification associated with those two records is indicated in green on the plot. The probability associated with the rare target's record is indicated in red. The rare target has a low disclosure risk both absolutely (nearly 0) and relative to the other two rare records as well as all of the records in the released SCHL/VPS category.

The bottom plot in Figure 4.27 displays the disclosure risk associated with all of the records in the same SCHL/VPS category as the common target. These points are shown in black. The common target has the same original values on SCHL and VPS as 112 other records, whose disclosure risk is plotted in red, and has the same SCHL/VPS category as 1,884 other records, whose disclosure risk is plotted in black. The common target's disclosure risk is plotted in green. The common target has a low disclosure risk of 0.00058 or 0.058% which is also low relative to the disclosure risk associated with other records in the same VPS/SCHL category. Again, there are a number of records with relatively high risk values which the agency may want to examine to determine if further perturbation is necessary before releasing the data set.

Naive intruder

The naive intruder is assumed to have no prior knowledge about the data set except for the target's values on available variables in Ad and Ap. We use the formulation in Section 3.2.4 for $C_{naive} = \prod \frac{1}{n_k}$, for record j when $t^{Ad} = z_j^{Ad}$ and $z_{kj}^{Ap} \in (t_k - \eta_k, t_k + \eta_k)$. We set $B_{naive} = 1$ and $A_{naive} = \frac{1}{\prod \phi_{kj}}$, for every j, where $\phi_{kj} = \phi\left(\frac{z_{kj} - \overline{z}_{kj}}{\sigma_k}\right)$. Disclosure risk associated with the naive intruder is presented in two lines of Table 4.11. In the line labeled *naive*, disclosure risk is 0 for each target. To compute this, C_{naive} is computed, with n_k set equal to the number of records where the released value z_{kj}^{Ap} is in the interval $(t_k - \eta_k, t_k + \eta_k)$ for records in the same SCHL/VPS category as the target record, and is set to 0 for other records. We set η_k equal to the standard deviation of released values on variable z_k^{Ap}, corresonding to AGE, RET, SSI, and TAX. For each of the targets, this was quite restrictive. In fact, none of the records in the targets' categories had values within one standard deviation of all four variables simultaneously. Thus, all records have disclosure risk 0 when risk is computed in this way.

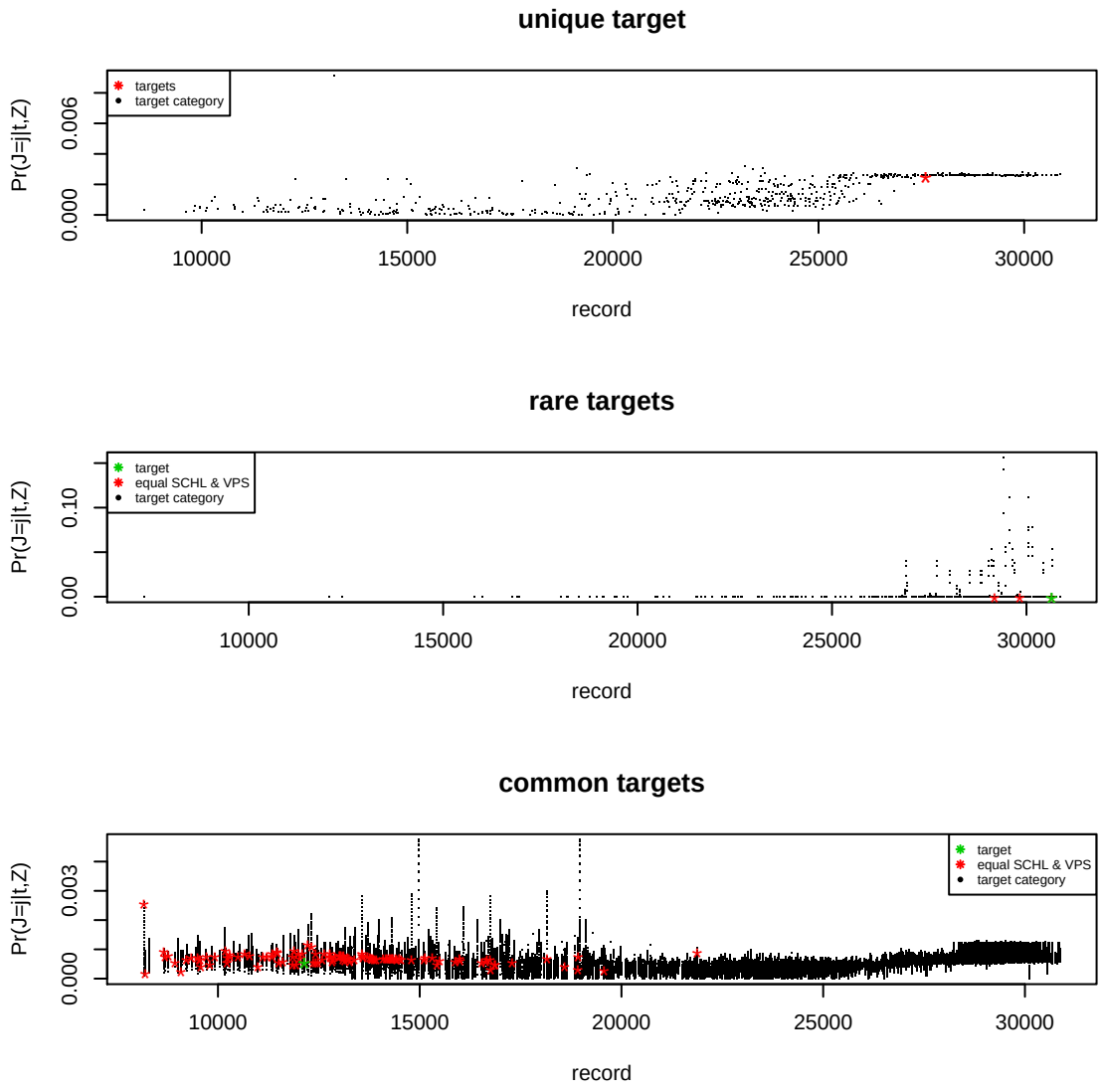

Figure 4.27 Risk for SDL intruder for records with SCHL/VPS value equal to the target.

Imagining that an intruder might take another approach, so do we. We determine that none of the records have a value of RET within one standard deviation of the common target's RET value. Instead of setting C_{naive} to 0, we loosen the restrictions and essentially set n_{RET} equal to 1. Then, component C_{naive} is computed using n_{AGE}, n_{SSI}, and n_{TAX} only. This produced the disclosure risk of 0.112, or 11.2% in Table 4.11 in the row labeled $naive^*$. Similarly, loosening restrictions for the unique and rare targets allowed records in their SCHL/VPS category to have non-zero disclosure risk as well. However, the synthetic records corresponding to each target did not contain values simulatneously within the intervals for remaining variables.

To assess the disclosure risk associated with the entire synthetic data set and associated with the target records relative other records, plots similar to those in Figure 4.27 can be examined. Many records have disclosure risk of 0 though, so such a plot is not very informative. Further investigation into the disclosure risk associated with 112 records with the same original SCHL and VPS values show that the minimum risk for any of these records is computed to be 0 and the maximum risk is 0.2586. This indicates that the common target's record is not identified with the highest disclosure risk among these records. Thus, while a value of 0.112 seems like a fairly high value for disclosure risk. Relative to other records in the common target's category, it is less than half of the risk calculated for at least one other record.

Average intruder

The average intruder possesses some knowledge of the data set prior to its release, though no specific information about the SDL procedure. In this application, we assume that parameter estimates from the following regression models are provided to the intruder.

$$\begin{aligned}
Y_{AGE} &= Y_{SCHL/VPS}\beta_{AGE} + \epsilon_{AGE} \\
logY_{RET} &= Y_{SCHL/VPS}\beta_{RET} + \epsilon_{RET} \\
Y_{SSI} &= Y_{SCHL/VPS}\beta_{SSI} + \epsilon_{SSI} \\
Y_{TAX} &= Y_{SCHL/VPS}\beta_{TAX} + \epsilon_{TAX} \\
logY_{SS} &= (Y_{SCHL/VPS}\ Y_{AGE}\ logY_{RET})\beta_{SS} + \epsilon_{SS} \\
logY_{WAGE} &= (Y_{SCHL/VPS}\ Y_{AGE}\ logY_{RET})\beta_{WAGE} + \epsilon_{WAGE}
\end{aligned}$$

Further, we assume that the intruder uses this information to compute the probability of

identification for each target in the released data set. The components are formulated as in Section 3.2.5, Equation 3.38, 3.41, and 3.42. Disclosure risk associated with the average intruder is computed to be 0 for each target, as presented in Table 4.11.

While values released in the synthetic versions of the target records, may be close to predicted values based on estimates from the above models, this does not occur for all six variables simultaneously. In fact, predicted values on one or more variables are quite far from the target's original values. The result is a very small probability associated with these variables. This small (nearly zero) value is multiplied with the probabilities associated with the other variables, leading to a very small (nearly zero) probability for the entire record. This occurs for all three targets under assumptions for the average intruder.

It should be noted that the models above do not necessarily represent conditional relationships in this data set well. It would be in the agency's best interest to consider what type of good analytical information will be available to the intruder and how s/he could incorporate that information into the probability calculation in an attempt to identify the target.

To assess disclosure risk in the entire data set under assumptions of the average intruder, similar plots to those in Figure 4.27 can be examined. Since all records with the same SCHL/VPS value as the target's have disclosure risk at almost zero, the plots are rather uninformative. Components of this risk associated with variables AGE, RET, SSI, and TAX are nearly zero for all records in each target's VPS/SCHL category. This implies that component C_{avg} is zero for each record. A small number of records contained non-zero probabilities for SS and WAGE, producing non-zero values of B_{avg} for some records. Further investigation into this could be informative.

CHAPTER 5. SUMMARY AND DISCUSSION

This chapter summarizes the work presented in this dissertation. Comments and suggestions for future work are included. Discussion on the proposed synthetic data method, disclosure risk measures, and data utility of the resulting synthetic data set are presented in this chapter.

5.1 Proposed method

In this dissertation an approach to generating synthetic data for statistical disclosure limitation is proposed. The method combines quantile regression, hot deck imputation, and rank swapping. Quantile regression estimates are used to compute predicted values for some sensitive variables. Hot deck imputation is used to identify records in the original data set with values close to the quantile regression predictions. From the closest record, values on additional variables are identified. Before imputing these values into the synthetic record, rank swapping is done to further perturb the values. The method is designed to produce a data set with high inferential validity and low disclosure risk. This combination of methods represents a unique approach to generating synthetic data for data sets with diverse variable types.

5.2 Disclosure risk measures

In order to assess the proposed method and to ensure confidentiality of respondents in the released synthetic data, disclosure risk in the resulting synthetic data set must be measured. Disclosure risk is dependent on three main components: the statistical disclosure limitation (SDL) method and resulting synthetic data, the intruder, and the records that might be targeted for identification. The extent to which the SDL method perturbs the data directly influences the disclosure risk. At two extremes, if data remain unperturbed they have high risk of disclosure. If data are perturbed beyond recognition, their disclosure risk may be close to zero, but their utility is likely too low. An agency that wishes to protect

sensitive data can assume an intruder will attempt to identify a target respondent in the released data. The agency should account for various assumptions pertaining to an intruder's knowledge about the SDL method used, knowledge about the data both before and after its release, and methods the intruder will use to identify respondents in the released data set.

In this dissertation, methods to measure disclosure risk are developed to address a range of intruders. The intruder with insider knowledge may have accurate information about the SDL method used and can decide to use that information to identify a target in the released data set. Working under a disclosure risk framework developed in Duncan and Lambert (1986, 1989) and Reiter (2005), we have extended risk measures that were developed to evaluate the situation when SDL methods such as noise addition, swapping, and recoding are used. Our extensions include a formulation of disclosure risk when the SDL method is to generate synthetic values using quantile regression, hotdeck, and rank swapping.

We present details on formulating disclosure risk when the intruder possesses no prior knowledge except for some of the target's original values. We call this the naive intruder and assume the such an intruder will base the probability of identifying a target in the released data on information gained after data are released only.

We also present a possible formulation for an average intruder with knowledge between that of the SDL and naive intruders. The average intruder's probability of identifying a target in the released data set is based on combining information available prior to the data release with information gained after the data are released. The assumptions about this intruder's knowledge and decisions are flexible to span the range between the naive and SDL intruders. Details of computing disclosure risk developed for these intruders for a data set generated using our proposed SDL method are a novel contribution to this area of study.

5.3 Data utility

In order to assess the utility of synthetic data generated using the SDL procedure proposed, we use a set of standard tools. Helpful visual tools include empirical plots of cumulative distributions and densities. We also consider comparisons of regression analysis. To compare results from several regression analyses, we present tables with parameter estimates and standard errors and using plots of standardized intervals.

5.4 Applications

The proposed SDL procedure has been applied in three settings. The first was at the Iowa Department of Revenue. There the purpose was to protect income tax return data for release to legislative researchers who investigate tax law changes and their effects on the revenue for the state of Iowa. The second application was at the U.S. Census Bureau. The purpose here was to protect veterans data from the American Community Survey database for release to researchers interested in studying various characteristics about the population of veterans. In our third application, we apply the proposed method to a Public Use Microdata Sample available from the U.S. Census Bureau. The purpose of this application is primarily to implement the disclosure risk measures developed in Chapter 3 to measure risk in a synthetic data set that was generated using the SDL method.

5.5 Future work

Each area of this research can be enhanced and improved by further work and research. In general, the SDL method can be further developed for application to a particular data set, possibly with known purposes once released to researchers and the public. Perhaps a particular portion of one of the surveys collected by the U.S. Census Bureau is in high demand by researchers interested in particular characteristics of the database. It is plausible that the proposed SDL method can provide a tool to generate high quality synthetic data for researchers studying those characteristics. With known uses, the data utility can be measured against specific standards and the method can be molded to best suit the purpose of the resulting synthetic data set.

Over the course of this research, tools to reduce computational effort were learned and used to make the implementation of quantile regression on one hundred quantiles feasible in a large data set. Perhaps with more investigation, efficiency can be improved so that quantile regression can be done for quantiles on finer divisions of the (0,1) interval.

Suggestions from other researchers in synthetic data have included ideas to better maintain conditional relationships in the data by selecting quantiles in an alternative way. Rather than selecting quantiles randomly for each record and each variable independently, perhaps modeling the quantiles could provide more consistency with the original data set.

Finally, one could investigate the use of ideas from multiple imputation with this method

of generating synthetic data. Theoretically, users with access to multiple replicates of a data set could better quantify uncertainty associated with inferences.

BIBLIOGRAPHY

Bassett, G. and Koenker, R. (1978). Asymptotic Theory of Least Absolute Error Regression, *Journal of the American Statistical Association.* **73 363** 618-622.

Bassett,G. and Koenker,R. (1982). An Empirical Quantile Function for Linear Models with iid Errors. *Journal of the American Statistical Association.* **77 378** 407-415.

Buchinsky, M. (1998). Recent Andvances in Quantile Regression Models: A Practical Guideline for Empirical Research. *The Journal of Human Resources.* **33 1** 88-126.

Dalenius, T. and Reiss, S. P. (1978). Data swapping–A technique for disclosure control. Confidentiality in Surveys. *Report 31.* Department of Statistics, University of Stockholm, Stockholm, Sweden.

Dalenius, T. and Reiss, S. P. (1982). Data Swapping–A Technique for Disclosure Control. *Journal of Statistical Planning and Inference.* **6** 73-85.

Dandekar, R. A., Cohen, M., and Kirkendall, N. (2002). Sensitive Microdata Protection Using Latin Hybercube Sampling Technique. *Inference Control in Statistical Databases.* Domingo-Ferrer, J. (ed.). Springer-Verlag. New York.

Domingo-Ferrer, J. and Franconi, L., eds. (2006). *Privacy in Statistical Databases.* Lecture Notes in Computer Science **4302** Springer-Verlag Berlin Heidelberg.

Doyle, P., Lane, J. I., Theeuwes, J. M., Zayatz, L., eds. (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Elsevier Science BV. Amsterdam, The Netherlands.

Duncan, G. T. and Lambert, D. (1986). Disclosure-Limited Data Dissemination. *Journal of the American Statistical Association.* **86 393** 10-18.

Duncan, G. T. and Lambert, D. (1989). The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics.* **7 2** 207-217.

Duncan, G. T., Jabine, T. B., and de Wolf, V. A. (1993) *Private lives and public policies: confidentiality and accessibility of government statistics.* National Academy Press. Washington D.C.

Federal Committee on Statistical Methodology (1978). Report on Statistical Disclosure Limitation and Disclosure-Avoidance Techniques, FCSM Working Paper 2. http://www.fcsm.gov/working-papers/spwp2.html.

Federal Committee on Statistical Methodology (1994). Report on Statistical Disclosure Limitation Methodology, FCSM Working Paper 22.
http://www.fcsm.gov/working-papers/spwp22.html. Revised 2005.

Federal Committee on Statistical Methodology, Confidentiality and Data Access Committee. (2002). Restricted Access Procedures. Accessed online 2008.

Fienberg, S. and Willenborg, L. (1998). Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data. *Journal of Official Statistics.* Statistics Sweden. **14** 4.

Fitzenberger, B., Koenker, R., and Machado, J. A. F. eds. (2001). *Economic Applications of Quantile Regression.* Physica-Verlag. Heidelberg. New York.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2004). Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers. NISS Technical Report.

Hahn, J. (1997). Bayesian Bootstrap of the Quantile Regression Estimator: A Large Sample Study. *International Economic Review.* **38 4** 795-808.

Hawala, S. (2008). Unpublished/in progress. U.S. Census Bureau. Washington, D.C.

He, X., Jureckova, J., Koenker, R., and Portnoy, S. (1990). Tail Behavior of Regression Estimators and their Breakdown Points. *Econometrica.* **58 5** 1195-1214.

Huckett, J. C. (2006). Base File Description for the Iowa Department of Revenue. Technical Report. Iowa State University and Iowa Department of Revenue. Iowa.

Jabine, T. B. (1993). Statistical Disclosure Limitation Practices of United States Statistical Agencies. *Journal of Official Statistics.* Statistics Sweden. **9 2** 427-454.

Karr, A. F., Kohnen,C. N., Oganian, A., Reiter, J. P., and Sanil A. P. (2006) A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician.* **60 3** 1-9.

Kinney, S and Reiter, J. (2007) Making public use, synthetic files of the Longitudinal Business Database. *JSM Proceedings, Government Statistics Section [CD-ROM].* Alexandria, VA: American Statistical Association.

Koenker, R. (1986). Strong Consistency of Regression Quantiles and Related Empirical Processes. *Econometric Theory.* **2** 191-201.

Koenker, R. (1988). Asymptotic Theory and Econometric Practice. *Journal of Applied Econometrics.* **3 2** 139-147.

Koenker, R. (2000). Galton, Edgeworth, Frisch and prospects for quantile regression in econometrics. *Journal of Econometrics.* **95** 347-374.

Koenker, R. (2002). *Quantile Regression.* Econometric Society Monograph Series, Cambridge University Press. New York, New York.

Koenker, R. (2004). Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis* **91 1** 74-89.

Koenker, R. and Bassett,G. (1978). Regression Quantiles. *Econometrica.* **46 1** 33-50.

Koenker, R., Bassett, G. (1982). Tests of Linear Hypotheses and L1 Estimation. *Econometrica.* **50** 1577-1583.

Koenker, R. and d'Orey (1987, 1994). Computing regression quantiles. *Applied Statistics.* **36** 383393 and **43** 410414.

Koenker, R. and Geling, O. (2001). Reappraising Medfly Longevity: A Quantile Regression Approach. *Journal of the American Statistical Association.* **96 454** 458-468.

Koenker, R. and Hallock, K. F. (2001). Quantile Regression: An Introduction. *Journal of Economic Perspectives.* **15 4** 143-156.

Koenker, R. and Hendricks, W. (1992). Hierarchical Spline Models for Conditional Quantiles and the Demand for Electricity. *Journal of the American Statistical Association.* **87** 58-69.

Koenker, R. and Machado, J. A. F. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association.* **94 448** 1296-1310.

Koenker, R., Ng, P., and S. Portnoy, S. (1994). Quantile Smoothing Splines. *Biometrika.* **81 4** 673-680.

Koenker, R. and Park, B.J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics.* **71** 265-285.

Koenker, R. and Portnoy, S. (1987). L-Estimation for the Linear Model. *Journal of the American Statistical Association.* **82 399** 851-857.

Koenker, R. and Xiao, Z. (2002). Inference on the Quantile Regression Process. *Econometrica* **70 4** 15831612.

Koenker, R. and Xiao, Z. (2004). Unit Root Quantile Autoregression Inference. *Journal of American Statistical Association.* **94** 775-787.

Koenker, R. and Xiao, Z. (2006). Quantile Autoregression. *Journal of American Statistics Association, with discussion and rejoinder.* **475** 980-1006.

Koenker, R. and Zhao, Q. (1994). L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics.* **3** 223-235.

Koenker, R. and Zhao, Q. (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory* **12 5** 793-813.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd edition.* New York: John Wiley.

Ma, L. and Koenker, R. (2006). Quantile Regression Methods for Recursive Structural Equation Models. *Journal of Econometrics.* **134** 471-506.

Machado, Jose A. F. and Silva, J. M. C. Santos (2005). Quantiles for Counts. *Journal of the American Statistical Association.* **100** 1226-1237.

Moore, R. (1996). Controlled Data Swapping Techniques For Masking Public Use Data Sets. Research Report **96/04.** U. S. Census Bureau, Statistical Research Division. Washington D.C.

Muralidhar, K. and Sarathy, R. (2002). Application of the Two-Step Data Shuffle to the 1993 AHS Data: A Report on the Feasibility of Applying Data Shuffling for Microdata Release. Research report prepared for U.S. Census Bureau. http://gatton.uky.edu/faculty/muralidhar/maskingpapers/

Muralidhar, K. and Sarathy, R. (2006). Data Shuffling–A New Masking Approach for Numerical Data. *Management Science.* **52 5** 658-670.

Nin, J., Herranz, J., and Torra, V. (2008). Rethinking rank swapping to decrease disclosure risk. *Data & Knowledge Engineering.* **64 1** 346-364.

Portnoy, S. and Koenker, R. (1989). Adaptive L-Estimation for Linear Models. *Annals of Statistics.* **17 1** 362-381.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacean Tortoise: Computability of Squared-error vs Absolute Error Estimators. *Statistical Science.* **12 4** 279-296.

Reiter, J. P. (2004). New Approaches to Data Dissemination: A Glimpse into the Future (?). *Chance.* **17 3** 12-16.

Reiss, S. P. (1984). Practical Data Swapping: The First Steps. *ACM Transactions on Database Systems.* **9 1** 20-37.

Rodriguez, R. and Hawala, S. (2006). Disclosure Avoidance for the American Community Survey Public-Use Microdata Samples: A Model-Based Approach for Group Quarters Data. Unpublished research report from 2006 Joint Statistical Meetings for U.S. Census Bureau. Washington, D.C.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics.* **9** 462-468.

Sanil, A., Gomatam, S., and Karr, A. F. (2003). NISS WebSwap: A Web Service for Data Swapping. *Journal of Statistical Software.* **8** 7.

Shlomo, N. and Young, C. (2006) Statistical Disclosure Control Methods Through a Risk-Utility Framework. *Lecture Notes in Computer Science.* Domingo-Ferrer and Franconi (eds.) **4302** 68-81.

Schlörer, J. (1981). Security of statistical databases: multidimensional transformation. *ACM Transactions on Database Systems.* **6 1** 95-112.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley. New York. (p. 80).

Statistics Canada Research Data Centres Guide for Researchers under Agreement with Statistics Canada (2005). http://www.statcan.ca/english/rdc/index.htm. Accessed 2008.

Takemura, A. (2002). Local Recoding and Record Swapping by Maximum Weight Matching for Disclosure Control of Microdata Sets. *Journal of Official Statistics.* **18 2** 275-289.

Taylor, J. W. and Bunn, D. W. (1999). A Quantile Regression Approach to Generating Prediction Intervals. *Journal of Applied Econometrics.* **45 2** 225-237.

United Nations (2005). Fundamental Principles of Official Statistics and Principles Governing International Statistical Activities. http://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.htm. Accessed 2008.

U.S. Census Bureau (2006). *Design and Methodology.* American Community Survey U.S. Government Printing Office, Washington, D.C. http://www.census.gov/acs/www/Downloads/tp67.pdf. Accessed April 2008.

U.S. Census Bureau (2007). 2006 American Community Survey Data Users Handbook. http://www.census.gov/acs/www/Downloads/Handbook2006.pdf. Accessed April 2008.

U.S. Census Bureau (2008). Data Protection and Privacy Policy. www.census.gov/main/www/aboutus.html. Accessed March 2008.

U.S. Centers for Disease Control (2002). National Center for Health Statistics Policy on Micro-data Dissemination. http://www.cdc.gov/nchs/data/NCHS Micro-Data Release Policy 4-02A.pdf. Accessed January 2008.

Virtual Research Data Center (2008). Cornell University. http://www.vrdc.cornell.edu/news/. Accessed 2008.

Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice.* Springer, New York.

Willenborg, L. and de Waal, T. (2001). *Statistical Disclosure Control in Practice, Second Edition.* Springer. New York.

Winkler, W. (2006). Modeling and Quality of Masked Microdata. *Statistical Research Division Research Report Series.* **RR2006-1** U.S. Census Bureau. Washington, D.C.

Zayatz, L. (2005). Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. Statistical Research Division, U.S. Census Bureau. Washington, D.C.

Zayatz, L. (2007) Disclosure Avoidance Practices and Research at the U.S. Census Bureau: An Update. *Journal of Official Statistics.* **23 2** 253-265.

Zayatz, L. (2007). Imputation for Disclosure. Statistical Research Division, U.S. Census Bureau. Washington, D.C.

Zhou, K. Q. and Portnoy, S. L. (1996). Direct Use of Regression Quantiles for Construct Confidence Sets in Linear Models. *The Annals of Statistics.* **24 1** 287-306.